Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Lehrstuhl für Genomorientierte Bioinformatik

Genomassemblierung komplexer Gräsergenome aus hoch-heterogenen Datensets.

Thomas Nussbaumer

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften genehmigten

Dissertation.

Vorsitzender:	Univ.-Prof. C. Schwechheimer
Prüfer der Dissertation:	
	1. Univ.-Prof. Dr. H.-W. Mewes
	2. Univ.-Prof. Dr. Chr.-C. Schön

Die Dissertation wurde am 13. Oktober 2014 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 31. Mai 2015 angenommen.

Abstract

Plants show huge differences in their genome sizes: Floating bladderwort (*Utricularia gibba*) (82 Mb) represents one of the smallest plant genomes whereas barley (*Hordeum vulgare* L. subsp. *vulgare*) with 5.1 Gb and bread wheat (*Triticum aestivum* L.) with even 17 Gb have dramatically larger genomes. The functional gene space is comparable between plants and ranges between approximately 20,000 to 40,000 genes for a diploid genome. For hexaploid bread wheat approximately 100,000 genes are estimated. The differences in genome size are mainly caused by repetitive elements and genome duplications. The large fraction of repetitive DNA of large plant genomes was so far a major roadblock to assemble larger plant genomes from next-generation sequencing (NGS) data. Therefore, new and alternative methods are necessary to anchor genes and to develop a genome template for these highly complex genomes, for which current NGS and assembly technologies fail to assemble entire chromosomes. Genome assemblies of such large complex genomes still require an anchored physical map as intermediate step towards a completely sequenced and assembled genome. In this thesis, I describe several methods that supported the establishment of an anchored and sequence/gene-enriched physical map for barley. The approach can be easily applied to other highly complex genomes with sparse sequences and abundant marker maps, shotgun sequencing data and a physical map. To make use of the full potential of a physical map, high density marker maps were used along with an exhaustive exploitation of synteny against fully sequenced and annotated reference genomes. During this study, different data resources became available and were integrated to form the "*gene-ome*" of barley. The *gene-ome* of barley is defined by densely anchored physical contigs with a genetic and physical position for the majority of genes. It uses high density marker maps and the *GBS* (*Genotyping-by-Sequencing*) technology. Synteny to closely related and much smaller reference genomes allowed an anchoring of additional sequence data. Overall, a high fraction

of genes was anchored to a genetic position and to the physical map. This was possible due to the sequencing of gene-rich clones and sequence information from both clone ends and NGS contigs. Synteny to fully sequenced reference genomes with complete transcriptome data and fully assembled chromosomes underpinned the high accuracy of anchored genes and further allowed gaining new insights into the correlation between genetic recombination and localization, gene evolution and interpretation of chromosomal segments potentially introgressed during breeding processes between different barley cultivars. To make use of the potential of an anchored physical map, the barley framework was also used to detect genes involved in disease resistance of various susceptible wheat accessions acting against *Fusarium graminearum* as well as a structural model for *Triticium aestivum* where similar approaches for anchoring sequence resources have been applied.

Zusammenfassung

Pflanzen variieren in ihren Genomgrößen: Das Genom des Zwerg-Wasserschlauchs (*Utricularia gibba*) zählt mit 82 Mb zu einem der kleinsten Pflanzengenome, während Gerste (*Hordeum vulgare* L. subsp. *vulgare*) mit 5.1 Gb und Weizen (*Triticum aestivum* L.) mit 17 Gb deutlich größere Genome aufweisen. Genomduplikationen und repetitive Elemente, die bis zu 90% des Genoms einnehmen sind die Hauptgründe für derartige Größenunterschiede. Innerhalb der Pflanzen ist die Genanzahl mit ungefähr 20,000-40,000 Genen für diploide Genome vergleichbar, für den hexaploiden Weizen werden 100,000 Gene angenommen. Aufgrund der vielen repetitiven Elemente können Sequenzen aus der Hochdurchsatz-Sequenzierung (NGS) mit derzeitigen Leselängen und Assemblierungsmethoden nicht zu vollständigen Chromosomen zusammengesetzt werden. Die daraus resultierenden fragmentarischen Teilassemblierungen erfordern Methoden, um ein Genomgerüst zu schaffen und die Anordnungen von Genen zu bestimmen. Dies erfordert die Zuhilfenahme einer verankerten physikalischen Karte als Zwischenschritt hin zu einem vollständig sequenzierten und assemblierten Genom. In dieser Arbeit beschreibe ich die Entwicklung von Methoden und Anwendungen, die es erlaubten, die physikalische Karte in Gerste trotz der zugrundeliegenden kurzen Sequenzen zu 80% zu verankern. Hier zeige ich dieses Vorgehen am Beispiel von Gerste und Weizen, es kann aber auch für andere, ähnlich komplexe Genome angewandt werden, sofern Markerkarten, Hochdurchsatz-Sequenzdaten (NGS) und eine physikalische Karte vorliegen. Um das Potential der physikalischen Karte voll auszuschöpfen, sind hochauflösende genetische Marker notwendig. Damit werden einige hundert Kilobasen (kb) große genomische Sequenzen ("fingerprinted contigs") aneinandergereiht. Schließlich ist die Zuhilfenahme von Syntenie zu vollständig assemblierten Genomen zur Verankerung von genomischen Bereichen ohne genetischer Marker erforderlich. Im Zuge dieser Arbeit wurden mehrere heterogene Datensätze als Gerste "*gene-ome*" zusammengefasst und schrittweise

um neue Datensätze ergänzt. Das *gene-ome* beschreibt die Anordnung eines Großteils der Gene. Die Verankerung wurde mit Hilfe von Syntenie zu nahe verwandten Genomen validiert. Das Gerste *gene-ome* wurde auch genutzt, um für Weizen aus Transkriptomdaten proteinkodierende Gene zu filtern und somit die Resistenz gegen *Fusarium graminearum* zu untersuchen. Das *gene-ome* diente auch als Vorlage für die Erstellung von physikalischen Karten einzelner Weizenchromosomen.

Acknowledgements

Die *PhD*-Arbeit ist in zwei Abschnitte unterteilt: Die Bestimmung des Gersten "*gene-ome*" zum einen, zum anderen die Anwendung der entwickelten Methoden auf andere Genome (*u. a.* auf das komplexere Weizengenom). Während der erste Teil meiner Arbeit die ersten beiden Jahre der *PhD*-Arbeit abdeckt, wurde in den letzten Jahren der Fokus der Analysen auf die Anwendung des *gene-ome* gelegt. Personen, die mich auf diesen beiden Abschnitten begleitet haben, sind alle Beteiligten des *IBSC* (Internationale Gerstengenom Sequenzierungskonsortium) sowie des *IWGSC* (Internationale Weizengenom Sequenzierungskonsortium). Bei der Auswertung der Weizenexpressionsdaten Personen der *BOKU* (Universität für Bodenkultur Wien). Im Rahmen der Verankerung der physikalischen Karte in Gerste sind vor allem Personen des *IPK* Gatersleben (Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung) zu nennen, deren Gespräche, Ratschläge und Konferenzen meine Arbeit geprägt und deutlich verbessern konnten: Dr. Ruvini Ariyadasa, Dr. Burkhard Steuernagel, Dr. Martin Mascher, Dr. Uwe Scholz sowie Dr. Nils Stein, dem Koordinator des *IBSC*. Besonderer Dank geht an Ruvini, Burkhard und Martin für wertvolle Hilfestellungen und für die gute Zusammenarbeit. Auf Seiten des *MIPS* sind folgende Personen zu nennen: Dr. Klaus Mayer, Manuel Spannagl, Michael Seidel, Dr. Matthias Pfeifer, Verena Prade, Dr. Heidrun Gundlach und Mihaela Martis. Im Rahmen des zweiten Abschnittes, der Anwendung der Daten zur Analyse der Resistenz gegen Pilzbefall und im Zuge der Analyse einzelner Weizenchromosomen geht mein Dank an Personen des *IPK* (Dr. Naser Poursarebani, Dr. Thorsten Schnurbusch) sowie *MIPS* (Dr. Karl Kugler, Dr. Kai Bader, Dr. Sapna Sharma) und *BOKU* (Dr. Hermann Bürstmayr, Dr. Wolfgang Schweiger, Christian Ametz und Dr. Gerald Siegwart) sowie an Personen der *Kansas State University* (Dr. Sunish Sehgal, Dr. Bikram S. Gill). Der größte Dank geht an meinen Gruppenleiter Dr. Klaus Mayer und an den Betreuer meiner *PhD*-Arbeit

Prof. Dr. Werner Mewes für die Möglichkeit, die *PhD*-Arbeit am *MIPS* machen zu dürfen. Ebenfalls Dank an Prof. Dr. Chris-Carolin Schön sowie Prof. Dr. Claus Schwechheimer. Besonderer Dank an Dr. Klaus Mayer für die Möglichkeit, in vielen Projekten mitzuarbeiten und den Hilfestellungen. Außerdem für die Hilfe während des Schreibprozesses, wertvollen Tipps, um die große Fülle an Informationen zu strukturieren. Allen Gruppenmitglieder besonderer Dank für Kritiken und die gute Zusammenarbeit.

Schließlich besonderer Dank an Dr. Georg Haberer und Dr. Wolfgang Schweiger für kritische Kommentare zur schriftlichen Arbeit sowie an das Prüfungskomitee für die Möglichkeit diese Arbeit deutlich zu verbessern.

Publikationsliste

Teile dieser Arbeit wurden veröffentlicht:

1. Ariyadasa*[1], Mascher*, **Nussbaumer***, Schulte, Frenkel, Poursarebani, Zhou, Steuernagel, Gundlach, Taudien, Felder, Platzer, Himmelbach, Schmutzer, Hedley, Muehlbauer, Scholz, Korol, Mayer, Waugh, Langridge, Graner, Stein (2014). A sequence-ready physical map of barley anchored genetically by two million single-nucleotide polymorphisms. Plant Physiol., 164, 1:412-23.

Beschreibung: Diese Studie beschreibt die Erstellung einer physikalischen Karte in Gerste, um wichtige Methoden und Erkenntnisse vorzustellen, die in der Studie des gesamten Gerstengenoms nur teilweise abgedeckt werden konnten. Der Nutzen von populationsgenetischen Methoden für die Verbesserung der physikalischen Karte ist darin dargestellt.

Mein Anteil: Umsetzung von diversen bioinformatischen Analysen, um die Verteilung von Klonbibliotheken auf der physikalischen Karte zu untersuchen. Zusätzlich die Entwicklung von Visualisierungsmethoden, mit denen die physikalischen Contigs mit ihren verbunden Sequenzelementen dargestellt sind. Diese Methoden dienten auch dazu, um schimärische Contigs zu bestimmen und um Contigs aufzuspüren, die zusammengefasst werden können. Kosten in der Sequenzierung des minimalen überspannenden Pfads in Gerste MTP (*minimum tiling path*) werden dadurch gesenkt.

2. Poursarebani*, **Nussbaumer***, Šimková, Šafář, Witsenboer, van Oeveren, Doležel, Mayer, Stein, Schnurbusch (2014). Whole-genome profiling

[1] Gleichwertige Beiträge.

and shotgun sequencing delivers an anchored, gene-decorated, physical map assembly of bread wheat chromosome 6A. Plant J., 79, 2:334-47.

Beschreibung: Diese Studie beschreibt die verankerte physikalische Karte des Weizenchromosoms 6A. Die mit der WGP^{TM} (*whole genome profiling*)-Technologie erzeugten und sehr kurzen Sequenzen wurden mit Scaffolds aus Weizen und dessen Vorläufergenomen erweitert. Die Studie beschreibt außerdem eine äußerst umfangreiche Gegenüberstellung von zwei oft verwendeten Assemblierungsprogrammen für physikalische Karten, um eine besonders robuste physikalische Karte zu schaffen.

Mein Anteil: In dieser Studie führte ich die genetische Verankerung der physikalischen FP contigs durch und übertrug heterogene Datensätze aus publizierten Studien auf die physikalische Karte. Im Zuge der vergleichenden Analyse der beiden Assemblierungsstrategien wurden deskriptive Analysen und textuelle Beiträge beigesteuert, diskutiert und beschrieben.

3. Kugler, Siegwart, **Nussbaumer**, Ametz, Spannagl, Steiner, Lemmens, Mayer, Buerstmayr, Schweiger (2013). 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, 14:728.

Beschreibung: Diese Arbeit beschreibt die Transkriptomanalyse von unterschiedlich resistenten Weizenlinien gegen Pilzbefall. Die Studie repräsentiert eine der ersten Studien mit der Beschreibung einer RNA-seq-basierten Transkriptomanalyse für zwei wichtige Resistenzbereiche auf dem Weizengenom. Auf Basis von den Transkriptomdaten wurde ein Ko-expressionsnetzwerk erstellt. Dadurch konnten Gene mit ähnlichen Expressionsprofilen bestimmt und funktionell analysiert werden.

Mein Anteil: In dieser Studie wurde die Zuordnung ("mapping") der RNA-seq Daten auf die Referenzsequenz, die Bestimmung von differentiell exprimierten Genen als auch die Analyse von selektierten, stark unterschiedlich regulierten Genfamilien durchgeführt. Außerdem wurden Gerstengene mit den Weizengenen assoziiert,

um funktionelle Annotationen zuzuweisen.

4. Wang, Haberer, Gundlach, Gläßer, **Nussbaumer**, Luo, Lomsadze, Borodovsky, Kerstetter, Shanklin, Byrant, Mockler, Appenroth, Grimwood, Jenkins, Chow, Choi, Adam, Cao, Fuchs, Schubert, Rokhsar, Schmutz, Michael, Mayer, Messing (2014). The Spirodela polyrhiza genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. Nat Commun, 5:3311.

Beschreibung: Die Studie beschreibt das Teichlinsengenom mitsamt dessen funktionellen und sequenzbasierten Annotationen. Das Genom stellt ein besonders interessantes Beispiel in der Pflanzenwelt dar, weil es im Vergleich zu anderen Pflanzen ein äußerst reduziertes Genom mit sehr geringer Genanzahl aufweist.

Mein Anteil: Die physikalische Karte wurde gegen die assemblierten NGS Scaffolds verglichen. Diese Überprüfung erfolgte durch Homologievergleiche und anschließender manuellen Überprüfung jedes Scaffolds und FP contigs, indem die physikalischen Distanzen auf den entsprechenden NGS Scaffolds gegen die Anzahl an Konsensusbanden verglichen wurde. Dadurch wurden rund 300 Scaffolds zu Chromosomen zusammengefasst. Außerdem wurden während des Evaluierungsprozesses vergleichende Analysen zu Referenzgenomen durchgeführt, um die Vollständigkeit des Transkriptoms zu überprüfen.

5. Mayer, Martis, Hedley, Simková, Liu, Morris, Steuernagel, Taudien, Roessner, Gundlach, Kubaláková, Suchánková, Murat, Felder, **Nussbaumer**, Graner, Salse, Endo, Sakai, Tanaka, Itoh, Sato, Platzer, Matsumoto, Scholz, Dolezel, Waugh, Stein (2011). Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell, 23, 4:1249-63.

Beschreibung: Diese Arbeit stellt ein erstes umfangreiches genomisches Modell für Gerste dar. Unter Zuhilfenahme der deutlich kleineren Referenzgenome von Reis, *Brachypodium* und *Sorghum* wurde eine virtuelle Verankerung von Gerstengenen unabhängig von einer physikalischen Karte vorgenommen.

Mein Anteil: Erstellung einer Methode, *chromoWIZ*, mit der die Bestimmung, Eingrenzung und Visualisierung von syntenischen Segmenten ermöglicht wird. Konstante Überarbeitungen des Programms wurden durchgeführt, um spezifische Aspekte dieser Publikation zu unterstützen. Dieses Programm ist ein zentrales Modul des GenomeZipper-Konzepts und wurde zunächst als Paket mit einzelnen Skripten in der Programmiersprache Python bereitgestellt, später als Webapplikation eigenständig publiziert (Nussbaumer et al. [2014b]).

6. Silvar, Perovic, **Nussbaumer**, Spannagl, Usadel, Casas, Igartua, Ordon (2013). Towards positional isolation of three quantitative trait loci conferring resistance to powdery mildew in two Spanish barley landraces. PLoS ONE, 8, 6:e67336.

Beschreibung: Beschreibung von drei spanischen Gersten QTLs mit Hilfe der zu diesem Zeitpunkt neuen physikalischen Karte in Gerste als Erweiterung zum Gersten GenomeZipper. Es erlaubte Gene zu bestimmen, die in darauf anschließenden Studien genauer analysiert werden können.

Mein Anteil: Suche nach FP contigs, die innerhalb von drei genetisch eingegrenzten Regionen liegen. Auflistung und funktionelle Beschreibung der Gene, die in diesem Intervall lokalisiert sind.

7. **Nussbaumer**, Kugler, Bader, Sharma, Seidel, Mayer (2014). RNASeqExpressionBrowser–a web interface to browse and visualize high-throughput expression data. Bioinformatics, 30, 17:2519-20.

Beschreibung: Ein Webportal, um die Suche und Analyse von Expressionsprofilen mit Hilfe von Sequenzen oder durch funktionelle Beschreibungen zu erleichtern. Der *RNASeqExpressionBrowser* wird lokal installiert und weist detaillierte Beschreibungen zu ausgewählten Genen auf.

Mein Anteil: Konzeption, Implementierung der Software und Erstellung des Manuskripts. Diese Ressource ist mittlerweile

eine zentrale Schnittstelle für unterschiedliche RNA-seq-basierte Projekte.

8. **Nussbaumer***, Martis*, Roessner*, Pfeifer, Bader, Sharma, Gundlach, Spannagl (2013). MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Res., 41, Database issue:D1144-51.

Beschreibung: Eine Beschreibung von Daten und Funktionalitäten, die durch MIPS PlantsDB seit Jahren bereitgestellt werden sowie neu entwickelt wurden. Die aktualisierte Darstellung berücksichtigt Entwicklungen innerhalb der letzten zwei Jahre.

Mein Anteil: Implementierung und Beschreibung von *CrowsNest*, einem Tool, das vergleichende Analysen zwischen Pflanzengenomen erlaubt (*Brachypodium distachyon*, *Oryza sativa*, *Hordeum vulgare*, *Triticum aestivum*). Berücksichtigung des Gerstengenoms in *CrowsNest*.

9. Spannagl, Martis, Pfeifer, **Nussbaumer**, Mayer (2013). Analysing complex Triticeae genomes - concepts and strategies. Plant Methods, 9, 1:35.

Beschreibung: Beschreibung von Konzepten, um Getreidegenome trotz vieler repetitiver Sequenzen zu analysieren. Diese Ansätzen umfassen den Gersten GenomeZipper, die Beschreibung des Weizen-Ortholoms, einem ersten genomischen Modell für Brot-Weizen sowie die Beschreibung der physikalischen Karte in Gerste.

Mein Anteil: Beschreibung der Relevanz und Erstellung der verankerten physikalischen Karte in Gerste.

10. Schweiger, Pasquet, **Nussbaumer**, Paris, Wiesenberger, Macadré, Ametz, Berthiller, Lemmens, Saindrenan, Mewes, Mayer, Dufresne, Adam (2013). Functional characterization of two clusters of Brachypodium distachyon UDP-glycosyltransferases encoding putative deoxynivalenol detoxification genes. Mol. Plant Microbe Interact., 26, 7:781-92.

Beschreibung: Die Studie beschreibt zwei funktionelle Gen-

gruppen in *Brachypodium* mit Sequenzhomologie zu funktionellen Genen aus *Arabidopsis* und Gerste, die verantwortlich für den Abbau von Pilztoxinen sind. Es handelt sich in dieser Studie vor allem um *in silico* Analysen.

Mein Anteil: Für einzelne Untergruppen der *Brachypodium* Gengruppen wurden evolutionäre Analysen durchgeführt.

11. Muñoz-Amatriaín, Eichten, Wicker, Richmond, Mascher, Steuernagel, Scholz, Ariyadasa, Spannagl, **Nussbaumer**, Mayer, Taudien, Platzer, Jeddeloh, Springer, Muehlbauer, Stein (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol., 14, 6:R58.

Beschreibung: In diesem Ansatz wurde die Anwendung einer Array-Hybridisierung zur Bestimmung der Häufigkeit von sehr kurzen genomischen Stücken in 14 Gerstenkultivaren analysiert. Diese Arbeit stellt eine der ersten Studien dar, die das Gersten *gene-ome* nutzen konnte und dessen Relevanz und Bedeutung zur Bestimmung von genetischer und genomischer Variabilität beleuchten konnte.

Mein Anteil: Die verankerte physikalische Karte wurde erstellt und bereitgestellt.

12. A physical, genetic and functional sequence assembly of the barley genome. International Barley Genome Sequencing Consortium. Nature, 491, 711–716.

Arbeitsbereich: Integration of physical/genetic map (**Nussbaumer** *et al.*).

Beschreibung: Die Studie beschreibt die verankerte physikalische Karte von Gerste mit allen darauf annotierten Gerstengenen und NGS contigs und stellt das erste in diesen Umfang vorliegende Getreidegenom dar, mit einer Größe von mehr als 5 Gb. Diese Arbeit wurde durch internationale Anstrengungen unterschiedlicher Institute ermöglicht.

Mein Anteil: Hauptverantwortlicher für das Arbeitspaket "Integration of physical/genetic map and sequence resources". Diese Arbeit umfasst die Integration von unterschiedlichen genetischen Markern, die Erweiterung durch NGS contigs sowie die Zuweisung von Genen und stellt den zentralen Teil dieser Promotionsarbeit dar.

13. **Nussbaumer***, Kugler*, Schweiger, Bader, Gundlach, Spannagl, Poursarebani, Pfeifer, Mayer (2014). chromoWIZ: a web tool to query and visualize chromosome-anchored genes from cereal and model genomes. BMC Plant Biol., 14, 1:348.

Beschreibung: Eine Webapplikation, die eine Suche von Genen der Pflanzen *Brachypodium*, Reis, Weizen und Gerste erlaubt. Außerdem stellt *chromoWIZ* Möglichkeiten bereit, um Expressionsdaten als auch Listen mit differentiell exprimierten Genen als Filterkriterien zu nutzen.

Mein Anteil: Konzeption, Implementierung und Ausfertigung des Manuskripts.

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

Arbeitsbereich: Comparative analysis of diploid, tetraploid and hexaploid wheat: Pfeifer, Sandve, **Nussbaumer**, Bader, Choulet, Feuillet, Mayer.

Beschreibung: Beschreibung des Weizengenoms anhand eines NGS-basierten Datensatzes, aufgetrennt in 21 Chromosomenarme. Dieser Datensatz wurde zur Bestimmung von 124,201 Genen genutzt.

Mein Anteil: Bestimmung von Pseudogenen in den Subgenomen.

Teile dieser Arbeit wurden noch nicht veröffentlicht:

15. Sehgal *et al.* Beschreibung der Weizenchromosomen 1D, 4D und 6D.

Beschreibung: Die Studie beschreibt die genetisch verankerten physikalischen Contigs aus den Weizenchromosomen 1D, 4D und 6D und die Verwendung von Sequenzen aus Klonpools, um Sequenzen auf die physikalische Karte zu übertragen.

Mein Anteil: Genetische Verankerung der physikalischen FP contigs sowie Assemblierung und Dekonvolution der Sequenzen zu den Klonen dieser FP contig-Karte. Evaluierung der Qualität der physikalischen Karte sowie Beteiligung am Entwurf und der Anfertigung der Publikation.

16. Kugler, **Nussbaumer**, Warth, Sharma, Ametz, Simader, Parich, Lemmens, Schuhmacher, Krska, Buerstmayr, Mayer, Schweiger. Multilayered dissection of the molecular bread wheat (*Triticum aestivum*) response to *Fusarium graminearum*.

Beschreibung: Aktualisierung der Studie aus Kugler et al. [2013] unter Verwendung von Genmodellen aus IWGSC [2014].

Mein Anteil: Bestimmung der Expression von Weizengenen nach Pilzbefall. Bereitstellung eines Ko-expressionsnetzwerkes und Integration in den *RNASeqExpressionBrowser* (Nussbaumer et al. [2014b]). Mitarbeit an der Konzeption der Studie.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

BBH *best bidirectional hit*, beste bidirektionale Treffer.

BOKU Institut für Bodenkultur in Wien, `http://www.boku.ac.at/`.

CB *consensus band*, Konsensusbande.

BES *BAC end sequence*, sequenziertes Klonende.

CarmA *Chromosome arm Assignment*, Chromosomenarm-Zuordnung einer Sequenz.

cM *centimorgan*, Centimorgan.

DEG *differentially expressed gene*, signifikant unterschiedlich exprimiertes Gen.

HC *high-confidence gene*, Gen mit Sequenzhomologie zu einem Referenzgenom.

HICF *High-Information Content Fingerprinting*, Vorgehensweise, um Klone zu FP contig zu assemblieren.

IBSC *The International Barley Genome Sequencing Consortium*, Gerstengenom Sequenzierungskonsortium.

IPK Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung, `http://www.ipk-gatersleben.de/`.

Kansas State University Kansas State University, `http://www.k-state.edu/`.

LTR *Long terminal repeat*, LTR-Retrotransposon.

MIPS *Munich Information Center for Protein Sequences*, `http://mips.helmholtz-muenchen.de/`.

N50 Anzahl an nach Größe aufsteigend sortierter Contigs, ab der die kumulative Länge der Contigs die Hälfte der Gesamtlänge aller Contigs überschreitet.

IWGSC *International Wheat Genome Sequencing Consortium*, Weizengenom Sequenzierungskonsortium, `http://www.wheatgenome.org/`.

LC *low-confidence gene*, Gen, ohne Sequenzhomologie zu einem Referenzgenom.

GBS *Genotyping-by-Sequencing*, Genotypisierung durch Sequenzierung.

MTP *minimum tiling path*, Auswahl einer minimal überlappenden Anzahl an Klonen aller FP contigs als Vorlage zur Sequenzierung des Genoms.

NGS *Next-Generation Sequencing*, Hochdurchsatz-Sequenzierung.

Repeats Repetitive Elemente.

sBAC *sequenced BAC-clone*, Sequenzierter Klon.

UCR University of California Riverside, `http://plantbiology.ucr.edu`.

WCS *whole chromosome sorted sequences*, Chromosomenarm sortierte Sequenzen.

WGP™ *Whole Genome Profiling*, partieller Verdau von Klone durch Restriktionsenzyme in Verbindung mit Sequenzierung von kurzen, die Schnittstelle umschließende Sequenzen.

Kapitel 1

Einleitung

1.1 Bedeutung von Gerste und Weizen

Gerste (*Hordeum vulgare* L.) und Weizen (*Triticum aestivum* L.) gehören zur Familie der Süßgräser und zählen zusammen mit Reis (*Oryza sativa*) und Mais (*Zea mays*) zu den wirtschaftlich bedeutendsten Getreidearten. Wildgerste (*Hordeum spontaneum*) hat ihren Ursprung im Vorderen Orient und wurde als eine der ersten Pflanzen vor rund 10,000 Jahren von Menschen kultiviert (Badr et al. [2000]). Kulturgerste (*Hordeum vulgare* L.) wird in Sommer- und Wintergerste unterteilt: Während Wintergerste unter anderem als Futtermaterial eingesetzt wird, dient Sommergerste als Braugerste. Im Jahr 2013 wurden weltweit 144.8 Tonnen Gerste geerntet. Zu den größten Gersteproduzenten zählt Russland (15.4 t), gefolgt von Deutschland (10.3 t) und Frankreich (10.3 t) (FAOSTAT [2013]). Weizen zählt ebenfalls zu den Süßgräsern und wird in unterschiedlichen Sorten angebaut: Weichweizen (*Triticum aestivum* L.) und Hartweizen (*Triticum durum*) werden beispielsweise für Brot bzw. Pasta benötigt, andere Weizenarten werden als Futtermittel für Vieh verwendet. Weltweit werden 714 Tonnen Weizen angebaut. Die Volksrepublik China (122 t) ist größter Weizenproduzent, gefolgt von Indien (94 t) und den Vereinigten Staaten (58 t) (FAOSTAT [2013]). Weizen und Gerste werden in freier Natur von Schädlingen wie Pilzen, Bakterien, Insekten, Viren oder Fadenwürmern befallen, die zu deutlichen Ernteausfällen führen können. Bestimmte Sorten der beiden Getreidearten weisen starke Resistenzen auf oder zeigen eine besondere Anfälligkeit. Neben diesen Schädlingen sind abiotische Faktoren wie Dürre, Hitze und eine unterschiedliche Salzkonzentration im Nährboden von besonderer Bedeutung (Hossain et al. [2012]). Auf dem Weg zu einer Erschließung der Resistenzmechanismen

und für die Suche nach besonders resistenten oder toleranten Pflanzensorten ist die Bestimmung von längeren genomischen Sequenzen, idealerweise der Chromosomen, eine Notwendigkeit.

1.2 Pflanzengenome

Das Genom der Ackerschmalwand (*Arabidopsis thaliana*) ist das erste und annähernd vollständig erschlossene Pflanzengenom (AGI [2000]). Die kleine Genomgröße (125 Mb), die anspruchslose Aufzucht und Haltung mit geringem Platzbedarf und kurzer Generationszeit, die Vielzahl genetischer Ressourcen und der diploide Chromosomensatz im Gegensatz zu den polyploiden Genomen vieler anderer Pflanzen stellten wichtige Gründe für ihre Sequenzierung dar. 2002 folgte mit Reis (*Oryza sativa*) (Goff et al. [2002], Yu et al. [2002]) das erste Gräsergenom. Mit Aufkommen neuer Sequenziertechniken und einer deutlichen Senkung der Sequenzierungskosten (Liu et al. [2012]) folgten in immer schnellerer Abfolge weitere Genome. Im Jahr 2014 wurden 85 Pflanzengenome mit deutlichen Unterschieden in den Genomgrößen beschrieben (CoGePedia [2014]): Das Genom des Zwerg-Wasserschlauchs (*Utricularia gibba*) (Ibarra-Laclette et al. [2013]) zählt mit nur 82 Mb zu einem der kleinsten Genome, Gerste (*Hordeum vulgare*, 5.1 Gb) (IBSC [2012]) und Weizen (*Triticum aestivum*, 17 Gb) (Brenchley et al. [2012]) weisen hingegen deutlich größere Genome auf und stellen zurzeit die größten sequenzierten Pflanzengenome dar (Abbildung 1.1). Die Sequenzierung von Pflanzengenomen wird seit 2005/2006 maßgeblich von der Hochdurchsatz-Sequenzierung (NGS) bestimmt (Review in Egan et al. [2012]) und erlaubt inzwischen die routinemäßige Bestimmung kleinerer Genome oder Re-Sequenzierungen. In Pflanzengenomen wird die NGS-Sequenzierung in Verbindung mit der physikalischen Karte für die Bestimmung des Genoms oder in Kombination mit Sequenzierung nach Sanger (Diguistini et al. [2009]) genutzt. Für größere Nutzpflanzen wie Gerste und Weizen führt die Illumina-Sequenzierung nicht zu vollständigen Genomen, sondern generiert viele kleinere Konsensussequenzen (Jia et al. [2013], Ling et al. [2013]). Sequenziertechnologien der dritten Generation (*u.a.* Pacific Biosciences (PacBio)) erzeugen längere Einzelsequenzen. In Quail et al. [2012] wurden verschiedene NGS-Sequenziertechnologien wie Illumina MiSeq, Ion Torrent PGM, PacBio RS, Illumina GAIIx und Illumina HiSeq 2000 gegenübergestellt: Die größten Anschaffungskosten liegen mit rund $700k für den Illumina HiSeq 2000 sowie PacBio RS vor, am günstigsten

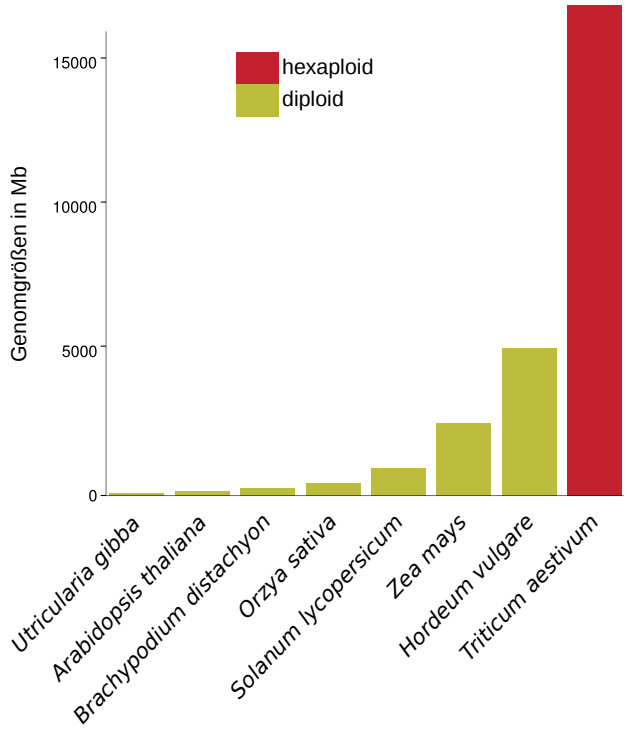

Abbildung 1.1: **Genomgrößen in ausgewählten Pflanzen**.
Beachtliche Genomgrößenunterschiede zwischen *Utricularia gibba* (Ibarra-Laclette et al. [2013]), *Arabidopsis thaliana* (AGI [2000]), *Brachypodium distachyon* (The International Brachypodium Initiative [2010]), *Oryza sativa* (Goff et al. [2002], Yu et al. [2002]), *Solanum lycopersicum* (Sato et al. [2012]), *Zea mays* (Schnable et al. [2009]), *Hordeum vulgare* (IBSC [2012]) und *Triticum aestivum* (IWGSC [2014]).

ist der Ion Torrent PGM ($80k). Die Sequenzlänglängen reichen von 700 bp für GAIIx/HiSeq 2000 bis zu 10 kb für die PacBio-Sequenzierung. Für PacBio-basierte Daten ist die Fehlerrate mit 12.9% deutlich höher als in den anderen Sequenziertechnologien (0.3-1.2% (HiSeq 2000 bzw. Ion Torrent)). Die größten Sequenzmengen pro Durchlauf werden von HiSeq 2000 (600 Gb) erreicht, gefolgt vom GAIIx (30 Gb). Mit PacBio und mit Ion Torrent können in einem Durchlauf maximal 1 Gb bzw. 100 Mb erzeugt werden. Die Längen einiger repetitiven Elementen von mehreren kb verhindern, dass Chromosomen vollständig zu einer durchgehenden Sequenz assembliert werden. In der Buche (*Picea glauca*) mit einer geschätzten Genomgröße von 20 Gb wurden die Illumina Sequenzierer HiSeq 2000 und MiSeq genutzt, und Insertgrößen von 250 bp, 500 bp, 6 kb, 8 kb und 12 kb wurden genutzt, um Einzelsequenzen zu Contigs und schließlich zu Scaffolds zusammenzufassen (Birol et al. [2013]). Die Assembly-Länge entspricht der erwarteten Genom-

größe, der *N50* beträgt 20 kb, in der Gerste beträgt der *N50* 2 kb.
Die hierarchische Sequenzierung und damit die Erstellung einer physikalischen Karte stellt derzeit immer noch die einzige Möglichkeit dar, um Chromosomen für sehr komplexe und hochrepetitive Genome wie Gerste und Weizen zu bestimmen (Steuernagel et al. [2009]).

1.2.1 Die physikalische Karte als Grundgerüst für die Genomsequenz

Eine "fingerprinted contig" (FP contig) Karte repräsentiert das Grundgerüst zur Bestimmung des Genoms (Abbildung 1.2). Zur Erstellung einer FP-Karte wird ein Genom zunächst von einem oder mehreren Restriktionsenzym(en) partiell geschnitten. Daraus entstehen einige hundert kb lange DNA-Fragmente, sogenannte Klone, die in einer BAC-Bibliothek kloniert und archiviert werden. Im nächsten Schritt, werden einzelne Klone durch Überlappungen ihrer Restriktionsmuster zusammengefasst. Ein minimaler überspannender Pfad (*minimum tiling path*, MTP) ist definiert als die Gruppe jener FP contigs, deren Summe die längste genomische Spannweite bei gleichzeitig geringster Anzahl von Klonen besitzt und stellt zugleich die effizienteste Vorlage für die Sequenzierung des Genoms dar. Nach Sequenzierung werden überlappende FP contigs zu Scaffolds und schließlich zu Chromosomen zusammengefasst. Alternativ werden Lücken durch genetische Karten oder Syntenie zu vollständig assemblierten Genomen geschlossen. Weizenchromosom 3B ist ein Beispiel, wie FP contigs in größeren Getreidearten zu einem Chromosom angeordnet wurden. Im Jahr 2008 wurde die genetisch verankerte physikalische Karte (Paux et al. [2008]) mit 1,036 FP contigs publiziert und im Jahr 2014 wurde das Chromosom unter großen finanziellen und personellen Anstrengungen bestimmt (Choulet et al. [2014]).

1.2.2 Genom-Sequenzierungen in der Gerste und in Weizen

Ein wesentliches Problem für die Assemblierung vieler Pflanzengenome stellt dessen außerordentlich hoher Anteil an repetitiven Sequenzen dar. Repetitive Regionen führen zu mehrdeutigen Pfaden und Verzweigungen in den Assembly-Graphen, die in der Regel nicht aufgelöst werden können und die Kontinuität der Contigs stark fragmentieren bzw. verkürzen (Übersicht in Alexeyenko et al. [2014]). Aufgrund der geringen Längen von NGS-Sequenzen von ≈100 bp für die Illumina Plattform und der deutlich größeren

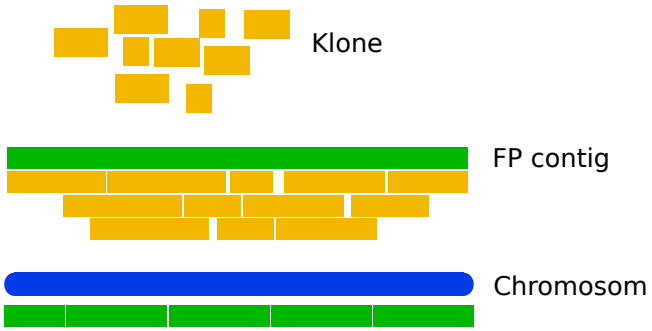

Abbildung 1.2: **Darstellung der hierarchischen Sequenzierung**.
Das Genom wird durch Restriktionsenzyme partiell verdaut. BAC-Klone werden zu FP contigs assembliert. Nach Sequenzierung der FP contigs entsteht das Chromosom.

Durchschnittslängen pflanzlicher Repeats von mehreren kb (*z. B.* 8.9 kb für *BARE-1* in Gerste, Shirasu et al. [2000]) werden bei (ausschließlicher) Verwendung von NGS-Sequenzen zahlreiche Brüche und eine geringe mittlere Contig-Länge erwartet. Das Verwenden von BAC-Klonen reduziert die Komplexität in der Assemblierung und erlaubt eine gezielte Sequenzierung eines 100 kb großen genomischen Bereichs (*u. a.* Steuernagel et al. [2009], Sato et al. [2011b]). In Steuernagel et al. [2009] wurde in der Gerste die Realisierbarkeit der Assemblierung von 91 BAC-Klonen mittels 454-Sequenzierung untersucht. Nach Assemblierung wurden ≈80% der BAC-Klone zu maximal 10 Teilsequenzen zusammengefasst, bei 24-facher Sequenziertiefe und einem *N50* von 63.69 kb. In Weizen wurde die Sequenz von dreizehn FP contigs mit einer Gesamtlänge von 18 Mb bestimmt (Choulet et al. [2010]). Für *Aegilops tauschii*, dem Vorläufergenom des Weizen D-Genoms, wurde eine verankerte physikalische Karte erstellt (Luo et al. [2013]). Hochdurchsatz-Sequenzierung erlaubte die Sequenzierung von Genomen im Tribus Triticeae, einer Pflanzengruppe mit vielen hoch-repetitiven Sequenzen. Für *Aegilops tauschii* (Ling et al. [2013]) und *Triticum urartu* (Jia et al. [2013]) wurden erfolgreich Genomsequenzen mit langer Kontinuität erstellt. Die Genomassemblierung von *Triticum urartu* lieferte 3.92 Gb der geschätzten Genomgröße von 4.94 Gb in Contigs. Nach dem "Scaffolding", dem Verbinden von Contigs durch Sequenzen mit größeren Insertionsgrößen, deckte das finale Assembly 4.66 Gb Sequenzen ab. Für *Aegilops tauschii* wurden mit variierenden Insertionsgrößen der Sequenzierbibliotheken bei 90-facher Genomabdeckung 83% des Genoms erschlossen und 43,150 proteinkodierende Genmodelle vorhergesagt. Für Weichweizen (*Triticum aestivum*; Brenchley

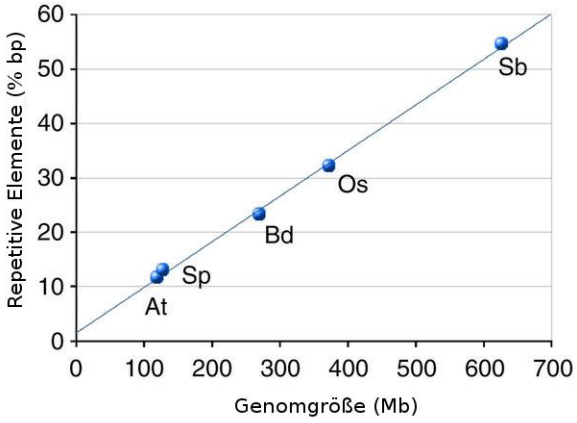

Abbildung 1.3: **Genomgröße und Anteil an repetitiven Elementen in Pflanzen**.
Vergleich der Genomgröße gegen den Anteil an repetitiven Elementen für ausgewählte Pflanzengenome (At=*Arabidopsis thaliana*, Sp=*Spirodela polyrhiza*, Bd=*Brachypodium distachyon*, Os=*Oryza sativa*, Sb=*Sorghum bicolor*). Für die angeführten Pflanzen mit Genomgrößen von <1 Gb kann aus dem Anteil an repetitiven Elementen auf die jeweilige Genomgröße geschlossen werden. Die Abbildung ist adaptiert nach Wang et al. [2014].

et al. [2012]) wurden auf Basis eines 454-Assembly ≈95,000 Genmodelle bestimmt. In IWGSC [2014] wurden einzelne Chromosomen und Subgenome bei rund ≈100-facher Genomabdeckung getrennt sequenziert und anschließend in einer ersten funktionellen Analyse 124,201 Gene annotiert. NGS contigs erlauben eine annähernd vollständige Erschließung des single- und low-copy Bereichs eines Genoms wie *z. B.* der kodierenden Sequenzen, aber nicht der Chromosomen.

1.2.3 Bedeutung von repetitiven Sequenzen für sequenzbasierte Genomanalysen

Hoch-repetitive Sequenzen erschweren massiv eine vollständige Erschließung der genomischen Struktur größerer Pflanzengenome. Der Gesamtanteil und die Frequenz der einzelnen Repeat-Klassen variieren in höheren Pflanzen sehr stark. Während in *Arabidopsis thaliana* nur rund 12-15% des Genoms durch wiederholende Basenpaarabfolgen, sogenannten "Repeats" eingenommen werden, sind es in der Gerste 80% (IBSC [2012]), bei Weizen sogar 90% (Brenchley et al. [2012]). Für ausgewählte Pflanzengenome ist der Anteil an repetitiven Elementen direkt proportional zur Genomgröße (Abbildung 1.3). In Mais weist die häufigste Repeat-Familie 52,000 Kopien auf,

während in *Arabidopsis* die höchste Kopienzahl einer Repeatklasse nur 194 beträgt. Repetitive Elemente werden in zwei Klassen unterteilt (Wicker et al. [2007]): Retrotransposons (Klasse 1) und DNA-Transposons (Klasse 2). Während Retrotransposons ihre Anzahl erhöhen, indem sie revers transkribiert werden und an anderen genomischen Bereichen ihre DNA-Sequenzen einfügen, ändern DNA-Transposons ihre Position und Zahl ohne einen RNA-Zwischenschritt entweder durch Exzision und Integration. Die Präsenz und Reihenfolgen von Domänen einzelner repetitiver Elemente erlauben eine feinere Gruppierung in Unterklassen. Die beiden größten Subklassen in den Retrotransposons sind *Gypsy* und *Copia LTR* (*Long terminal repeats*)-Retrotransposons. Wenige Repeat-Familien sind für diese Größenunterschiede verantwortlich (El Baidouri and Panaud [2013]): Bestimmte repetitive Elemente wie *BARE-1* decken in Gerste 12.7% des Genoms ab, vierzehn Familien von repetitiven Elementen sogar 50% (Wicker et al. [2009b]). In Reis (*Oryza brachyantha*, Chen et al. [2013]) werden 7.5% des Genoms durch sogenannte "*Mutator*-like repeats" eingenommen, die zu den DNA-Transposons zählen. Repetitive Elemente können über mehrere hundert kb ineinander verschachtelt sein und Insertionen von Repeats ineinander ("repeat junctions") als spezifische Marker zur Verankerung der physikalischen Karten genutzt werden (Paux et al. [2008]). Die Unterschiede in den Genomgrößen in den Pflanzen können durch diese repetitiven Elemente erklärt werden, die Anzahl und Anordnung von Genen ist hingegen stark konserviert.

1.3 Syntenie als Verankerungsstrategie

NGS-Assemblies in hoch-repetitiven Genomen erlauben die Bestimmung von Genen. Diese Assemblies sind aber nicht positionell angeordnet und liegen fragmentarisch vor (IBSC [2012], IWGSC [2014]). Analysen zur positionellen Klonierung von Zielgenen erfordern aber eine klare Bestimmung der genomischen Position eines Gens. Die vollständige chromosomale Sequenz erleichtert auch die Annotationen von Genen und erlaubt Strukturvergleiche zwischen Genomen. Vollständig assemblierte Genome bilden auch die Grundlage für Re-sequenzierungsansätze (Long et al. [2013]). In nahe verwandten Genomen weisen Gene eine ähnliche Genreihenfolge auf (Devos et al. [1995]). Mit fortlaufender evolutionärer Distanz wird der Grad der Konservierung geringer, da Genomduplikationen, Deletionen, Insertionen, aber auch Translokationen auf die genomischen Sequenzen und die Abfol-

ge der Gene Einfluss nehmen. Die Notwendigkeit, trotz technischer Grenzen, der durch Assemblies zu erreichenden Längen der Sequenzabschnitte, in Gerste die Anordnung von Genen zu ermöglichen, führte zur Entwicklung der GenomeZipper-Strategie. Der GenomeZipper ist ein ausschließlich auf Syntenie basierendes Verfahren, um Gene in einem Genom anzuordnen, für das keine durchgehenden chromosomalen Sequenzen vorliegen, sondern lediglich fragmentierte. Das Verfahren wurde erstmals in Gerste für einzelne, in Chromosomenarme sortierte Sequenzen mit niedriger Abdeckung eingesetzt, um Gene des Chromosoms 1 (1H) in Gerste linear anzuordnen (Mayer et al. [2009]).

Kleinere Genome wie Reis, *Brachypodium* und *Sorghum* wurden bereits zu Chromosomen assembliert und annotiert. Das Reisgenom (Goff et al. [2002], Yu et al. [2002]) wurde als Modellgenom genutzt, um andere nahe verwandte Gräser strukturell zu analysieren (Bolot et al. [2009]). Gerstengene sind über weite Strecken kollinear zu diesen Gräsern und werden in etwa 30 syntenische Blöcke eingeteilt (Bolot et al. [2009]). Eine klare Abgrenzung der Blöcke ist notwendig, um eine möglichst große Anzahl an Genen syntenisch anzuordnen. Die Durchflusszytometrie erlaubt für Gerste und für andere Getreidearten Chromosomen aufgrund von Deletions- und Substitutionslinien zytologisch zu trennen und zu sequenzieren (Vrana et al. [2000]). Zusammen mit genetischen Karten (Close et al. [2009], Sato et al. [2011a]) ermöglichte dies eine lineare Anordnung der syntenischen Bereiche. Im Gersten GenomeZipper wurden *Sorghum*, *Brachypodium* und Reis genutzt, syntenische Blöcke mit *chromoWIZ* (Nussbaumer et al. [2014b]) bestimmt und schließlich mit einer hoch-auflösenden genetischen Markerkarte (Close et al. [2009]) in eine lineare Anordnung gebracht. Neben Weizen und Roggen wurde die Methodik des Gersten GenomeZipper auch für das Deutsche Weidelgras (*Lolium perenne*) (Pfeifer et al. [2013]) eingesetzt, einem Genom mit einer sehr geringen Anzahl an Sequenzen und Markern. Für Weizen wurde eine ähnliche Art der Verankerung für Chromosom 4A (Hernandez et al. [2012]) genutzt. Es konnten mit diesem Modell 85% der Gene dieses Chromosoms in eine syntenische Anordnung gebracht. Abbildung 1.4 illustriert die Funktionsweise des GenomeZipper für Chromosom 5A in Weizen unter Zuhilfenahme von *Brachypodium*: Zunächst werden syntenische Bereiche definiert: Syntenische Bereiche liegen zu *Brachypodium* Chromosom 1 und Chromosom 4 vor, dargestellt durch *heat map* Darstellungen der Dichte an orthologem Gen. Anschließend erlauben Weizenmarker eine Ausrichtung dieser syntenischen Bereiche, wodurch das Pseudochromosom in

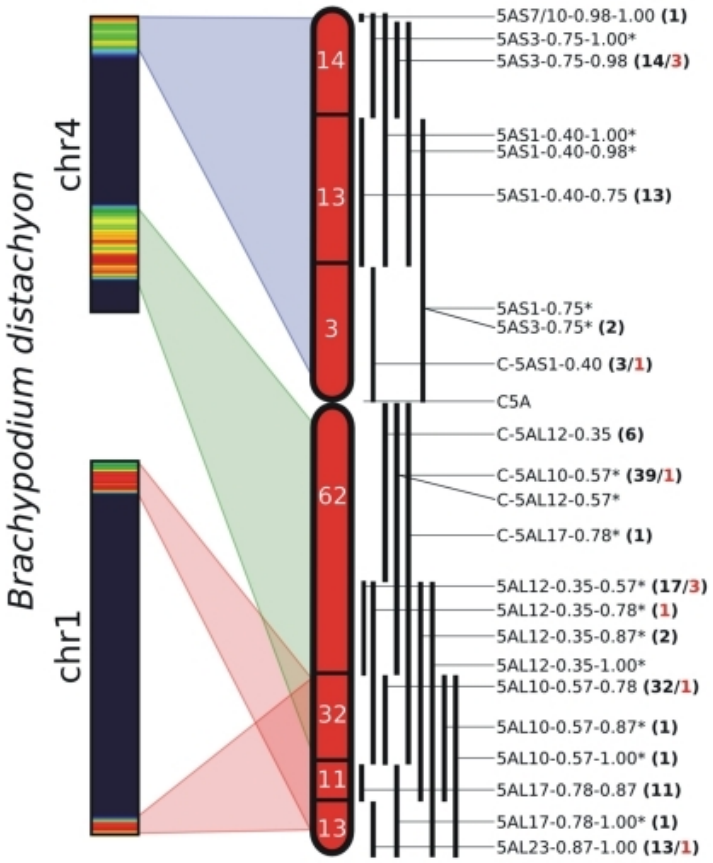

Abbildung 1.4: **Strategie des GenomeZipper**.
GenomeZipper Strategie für Chromosom 5A in Weizen. Syntenische Bereiche zwischen Weizenchromosom 5A und den Chromosomen in *Brachypodium* werden bestimmt. Anschließend werden genetische Marker genutzt, um die syntenischen Bereiche anzuordnen. Darstellung entnommen aus Vitulo et al. [2011].

Weizen entsteht.

1.4 Einfluss von Pilzbefall auf Weizen

Die Bestimmung von Referenzsequenzen ist eine Notwendigkeit, um besonders wichtige Nutzpflanzen zu untersuchen und damit Mechanismen zu erschließen, die der Pflanze helfen, mit biotischen und abiotischen Einflüssen umzugehen. Weizen, Gerste und viele andere Getreidearten sind einer Reihe unterschiedlicher Pathogene ausgesetzt. Pathogene wie Pilze (*u.a. Fusarium graminearum*) führen zu drastischen Ernteausfällen und schädigen Mensch

und Tiere, wenn die kontaminierte Pflanze oder Produkte aus dieser verzehrt werden. In Ährenfusariosen kommt es ungefähr 30-50 Stunden nach Infektion der Pflanzen zu einem deutlichen Anstieg an Toxinen (Mykotoxine, (*u.a.* Deoxynivalenol (DON)) und Tage später treten deutliche Verfärbungen der Blätter auf (Review in Wegulo [2012]). Pflanzen entwickeln aber Abwehrmechanismen und können Mykotoxine in weniger giftige Stoffe verwandeln (*u. a.* DON in DON-3-O-Glucosid (D3G), Karlovsky [2011]). Der unterschiedliche Befallsgrad zwischen einzelnen Arten resultiert aus der Quantität und Qualität von Resistenzbereichen. Für Ährenfusariosen ist die Weizensorte "Sumai-3" besonders resistent (Zhou et al. [2002]), andere sind besonders anfällig (*u. a.* "Remus") (Schweiger et al. [2013]). Für die Züchtung resistenter Sorten und zur Vermeidung von Einbußen durch Ernteverluste ist die Suche nach genomischen Bereichen, die für Resistenzen verantwortlich sind, von großer Bedeutung. Für die Resistenz gegen Ährenfusariosen wird das Gen *Fhb1* als jenes beschrieben, das sich hauptverantwortlich für die Resistenz zeigt. Der Genlokus wurde auf Weizenchromosom 3B genetisch eingegrenzt (Cuthbert et al. [2006]). Das Gen und ein weiterer Resistenzbereich (QTL) auf Weizenchromosom 5A (*Qfhs.ifa-5A*, Schweiger et al. [2013]) erklären den Großteil der Varianz zwischen resistenten und empfänglichen Weizenpflanzen. Die genetische Eingrenzung und funktionelle Analyse dieser Resistenzbereiche wird in dieser Arbeit analysiert.

1.5 Aufgabenstellung

Die Assemblierung und die Anordnung von Sequenzen aus der Hochdurchsatz-Sequenzierung für komplexere Pflanzengenome weisen in ihrer Anwendung Grenzen auf. Unterschiedliche Datensätze (sequenzierte Klone, NGS contigs, Marker) sollen idealerweise miteinander verbunden werden, um einen Datensatz zu generieren, der einem kompletten Genom/Chromosom möglichst nahe kommt. Folgende Ziele sollen erreicht werden:

1. Eine maximale Anzahl an FP contigs soll verankert werden.
2. Eine maximale Anzahl an Genen soll verankert werden.
3. Die Qualität der FP contig-Assemblierung soll überprüft werden.
4. Die Syntenie zu Referenzgenomen soll genutzt werden, um weitere FP contigs/Gene genetisch zu positionieren.

5. Heterogene Datensätze wie NGS contigs sollen integriert und interpretiert werden und ein Gerste *gene-ome* als zentrale Ressource der verankerten Gerstendaten soll geschaffen werden.
6. Strategien aus Gerste sollen auf Weizen übertragen werden (Chromosomen 1D, 4D, 6D und 6A).
7. Die neu entstandenen Ressourcen sollen genutzt werden, um zwei Resistenzbereiche gegen Pilzbefall auf Weizen zu untersuchen.

1.6 Struktur der Arbeit

In Abbildung 1.5 wird eine Übersicht der in dieser Arbeit integrierten Projekte und den Verbindungen zwischen den einzelnen Ressourcen gegeben. Das Gerste *gene-ome* (IBSC [2012]) stellt genetisch verankerte Gene und Ressourcen bereit, um später Genkandidaten aus einer Transkriptomstudie zur Resistenz gegen Ährenfusariosen (*Fusarium head blight*) in Weizen zu untersuchen (Kugler et al. [2013]) und mit umfangreicheren genomischen Daten neu zu interpretieren (Kugler et al. [In Vorbereitung]). Ansätze, die in der Gerste entwickelt wurden, werden anschließend auf das Weizenchromosom 6A (Poursarebani et al. [2014]) angewandt, der deutschen Beteiligung an der Sequenzierung des Weizengenoms. Methodiken aus der Gerstenverankerung werden für Weizenchromosomen 1D, 4D und 6D (Sehgal et al. [In Vorbereitung]) eingesetzt und Klone des minimal zusammengehörenden Pfads (*minimum tiling path*) werden assembliert. Die in dieser Arbeit entwickelten und publizierten Webtools *chromoWIZ* (Nussbaumer et al. [2014b]) und *RNASeqExpressionBrowser* (Nussbaumer et al. [2014a]) werden zur Interpretation der Transkriptomdaten aus Weizen genutzt.

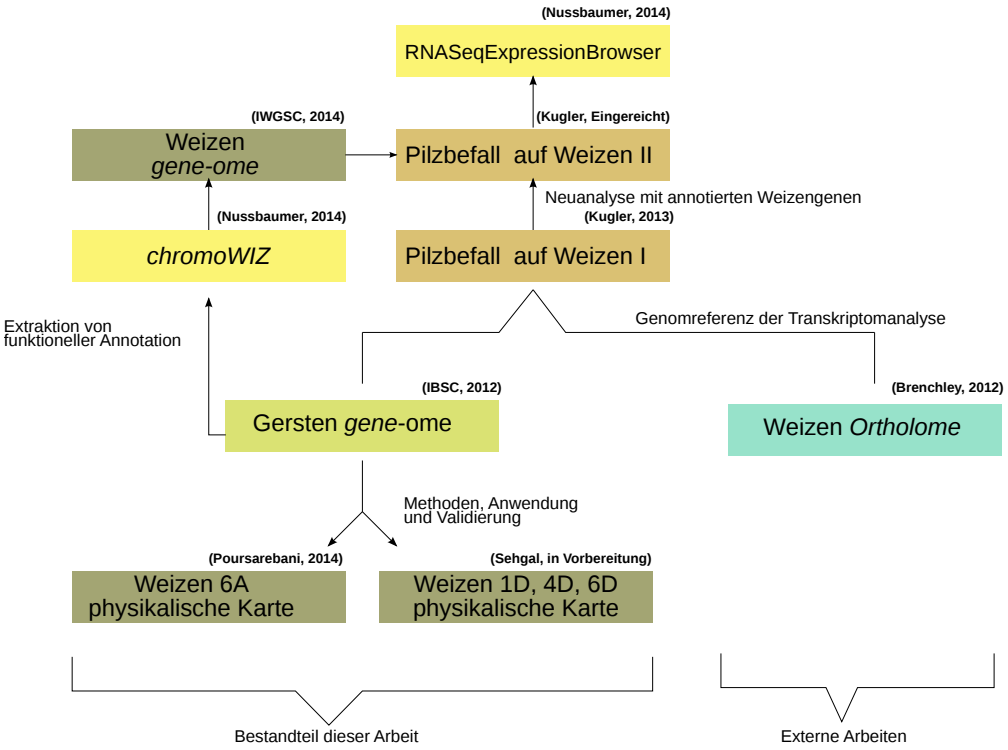

Abbildung 1.5: **Verbindungen der in dieser Arbeit beschriebenen Projekte**.
Das Gerste *gene-ome* stellt Ressourcen bereit, die in einzelnen Weizenchromosomen genutzt werden. Für die Analyse der Datensätze wurden Methoden entwickelt, um genomische und funktionelle Daten zu suchen.

Kapitel 2

Methoden

2.1 Erstellung der physikalischen Karten

Für die Erstellung der FPC-Karte in Gerste wurden sieben Klonbibliotheken mit vier unterschiedlichen Restriktionsenzymen (*RE*) (Tabelle 3.1) verwendet. Im Weizenchromosom 6A wurden zwei Klonbibliotheken verwendet und ein Restriktionsenzym (*Hin*dIII) genutzt, in Weizenchromosomen 1D, 4D, 6D drei Klonbibliotheken mit zwei unterschiedlichen Restriktionsenzymen (*Eco*R1, *Hin*dIII). Eine Klonbibliothek wird erstellt, indem hochmolekulare genomische DNA von einem Restriktionsenzym partiell geschnitten wird. Die erhaltenen genomischen DNA-Fragmente mit einer Länge von 100 - 200 kb werden anschließend in BAC-Vektoren kloniert und propagiert. Zur Erstellung des "fingerprint" werden die Klone vollständig von Restriktionsenzymen geschnitten, nach ihrer Größe auf einem Gel aufgetrennt und die entsprechenden Bandenmuster auf gemeinsame Teilmuster untersucht. In diesem Schritt werden überlappende Klone mit verfügbaren Programmen (LTC Frenkel et al. [2010], FPC Soderlund et al. [1997]) zu FP contigs zusammengefasst. Bei hoch-repetitiven Genomen wie Gerste und Weizen beginnt man mit einer sehr stringenten Klonassemblierung und reduziert die Stringenz kaskadenförmig (*z. B.* e^{-90} bis e^{-45} mit FPC in Gerste; e^{-75} bis e^{-11} mit LTC in Weizenchromosom 6A und e^{-75} bis e^{-10} mit FPC in Weizenchromosom 6A; derzeit e^{-60} mit FPC in Weizenchromosomen 1D, 4D und 6D). FP contig-Paare werden anschließend zusammengefasst, wenn genetische Marker auf beiden Enden des FP contigs die Überlappung bestätigen oder erst nach Sequenzierung des minimalen überspannenden Pfads. Mit der WGP$^{\text{TM}}$-Technologie (*Whole Genome Profiling*) werden Klone mit unterschiedlichen Restriktionsenzymen verdaut und Badenmuster mit LTC

(Frenkel et al. [2010]) und FPC (Soderlund et al. [1997]) assembliert sowie Sequenzen bereitgestellt, die eine Restriktionsschnittstelle umgeben. Die kurzen Sequenzen dienen in Folge als Sequenzanker für längere Sequenzen.

2.2 Marker

Marker werden in experimentelle und *in silico* Marker unterteilt. Experimentelle Marker werden durch Hybridisierungen von genetischen Markern auf Klone verankert, *in silico* Marker stellen Sequenzen bereit, die auf die verfügbaren sequenzierten Klone oder Klonenden über Sequenzhomologie zugewiesen werden. *GBS* (*Genotyping-by-Sequencing*) Marker sind *in silico* Marker. Diese Technologie reduziert die Genomkomplexität durch Auswahl bestimmter Restriktionsenzyme, die hoch-repetitive Regionen vermeiden. Die *GBS*-Technologie generiert zudem große Markermengen (*u. a.* 200,000 Marker in Mais und 25,000 Marker in Gerste (Elshire et al. [2011])). Aufgrund ihrer hohen Markerdichte und der gleichmäßigen Verteilung auf dem Genom sind *GBS*-basierte Marker vorteilhaft für die Verankerung von physikalischen Karten (*u. a.* Saintenac et al. [2013], Ariyadasa et al. [2014]). Experimentelle Marker werden durch Hybridisierungen (*u. a.* EST) auf Klone bestimmt. Experimentelle Marker sind präziser als *in silico* Marker, wenn wie in Gerste und Weizen ein Genom nur teilweise assembliert wurde und erlauben es, die Signifikanz eines Treffers auf Basis eines vollständigen Genoms zu untersuchen. In Gerste wurden experimentelle und *in silico* Marker genutzt (Tabelle 3.9), in den Weizenchromosomen 1D, 4D und 6D sowie 6A wurden nur *in silico* (Abbildung 3.12) Marker genutzt.

2.3 Integration von heterogenen Daten in Gerste

Vmatch (`www.vmatch.de`) erlaubte *in silico* Marker (Saintenac et al. [2013], Poland et al. [2012]) mit einer Mindesttrefferlänge von 55 bp und einer maximalen Hamming-Distanz von 1, den physikalischen Contigs mit allen darauf verankerten Ressourcen (sequenzierte Klone, Klonenden, Genomassemblierungsdatensätze) zuzuweisen. Die Mindestlänge von 55 bp wurde gewählt, um auch die kurzen Längen der *GBS* Marker von ≈60 bp zu berücksichtigen. Für die Erstellung der integrierten Konsensuskarte wurden alle *in silico* Markersequenzen genutzt, die einem FP contig zugewiesen wurden. Jeweils der Treffer mit dem höchsten Bitscore wurde weiter analysiert, bei gleichem Bitscore wurde der erste Treffer berücksichtigt. 113,117

Marker im Vergleich zu den 498,165 *in silico* Markern insgesamt wurden zusammen mit 3,276 experimentellen Markern als Basis für die Integration der Karten verwendet. Marker wurden entfernt, wenn eine Chromosomenzuordnung nicht der Chromosomenzuordnung des FP contigs auf Basis aller verankerten Sequenzen entsprach. Die Chromosomenzuordnung eines FP contigs wurde mit *CarmA* (siehe Kapitel 3.1.5) bestimmt. Verbliebene Markerzuweisungen wurden pro FP contig in ihre einzelnen Markerkarten gruppiert und jeweils der Median aller genetischen Positionen pro Karte bestimmt. Markerkarten, für die experimentelle als auch *in silico* Marker vorlagen, wie beispielsweise die SM3/MM3 Karten, wurden zusammengefasst, wenn dieselbe genetische Karte verwendet wurde. Schließlich wurde mit Hilfe der R-Funktion '*approxfun*' eine Funktion bestimmt, mit der die Länge der jeweiligen Karten auf die Länge der Referenzkarte normiert wurde. Anschließend wurden die einzelnen physikalischen Contigs nach ihrer genetischen Position angeordnet (Median der genetischen und nun SM6-normalisierten Positionen) und ihre physikalischen FP contig-Größen aufsummiert, wodurch die Gersten-Pseudochromosomen bestimmt werden konnten. Die Vorgehensweise ist in Abbildung 2.1 dargestellt. Dasselbe Vorgehen, das für Gerste beschrieben wurde auch zur genetischen Verankerung von NGS contigs sowie von Genen durchgeführt. Die Ergebnisse der Integration wurden online verfügbar gemacht (`ftp://ftpmips.helmholtz-muenchen.de/plants/barley/public_data/anchoring/`).

2.4 Integration von heterogenen Daten in Weizenchromosom 6A

Vmatch (`www.vmatch.de`) erlaubte die Zuordnung eines WCS contigs (bereitgestellt durch das IWGSC [2014]) zu physikalischen Contigs, wenn jeweils zwei WGP^{TM} *Tags* in entgegengesetzter Orientierung und getrennt durch die Erkennungssequenz des Restriktionsenzyms auf einem WCS verankert wurden. Fehler im Alignment zwischen WGP^{TM} *Tags* und WCS contig wurden nicht erlaubt.

Contigs aus den publizierten Studien der Weizenvorläufergenome *Triticum urartu* (Ling et al. [2013]) und *Aegilops tauschii* (Jia et al. [2013]) wurden genutzt, um WCS contigs zu erweitern, vorausgesetzt, eine Trefferlänge übertraf 200 bp bei einer Mindestidentität von 99%. Falls ein verankerter *T. urartu* contig Sequenzhomologien zu weiteren WCS contigs aufwies, wurden auch diese verankert. Die Verankerung der genetischen Karte in Weizen-

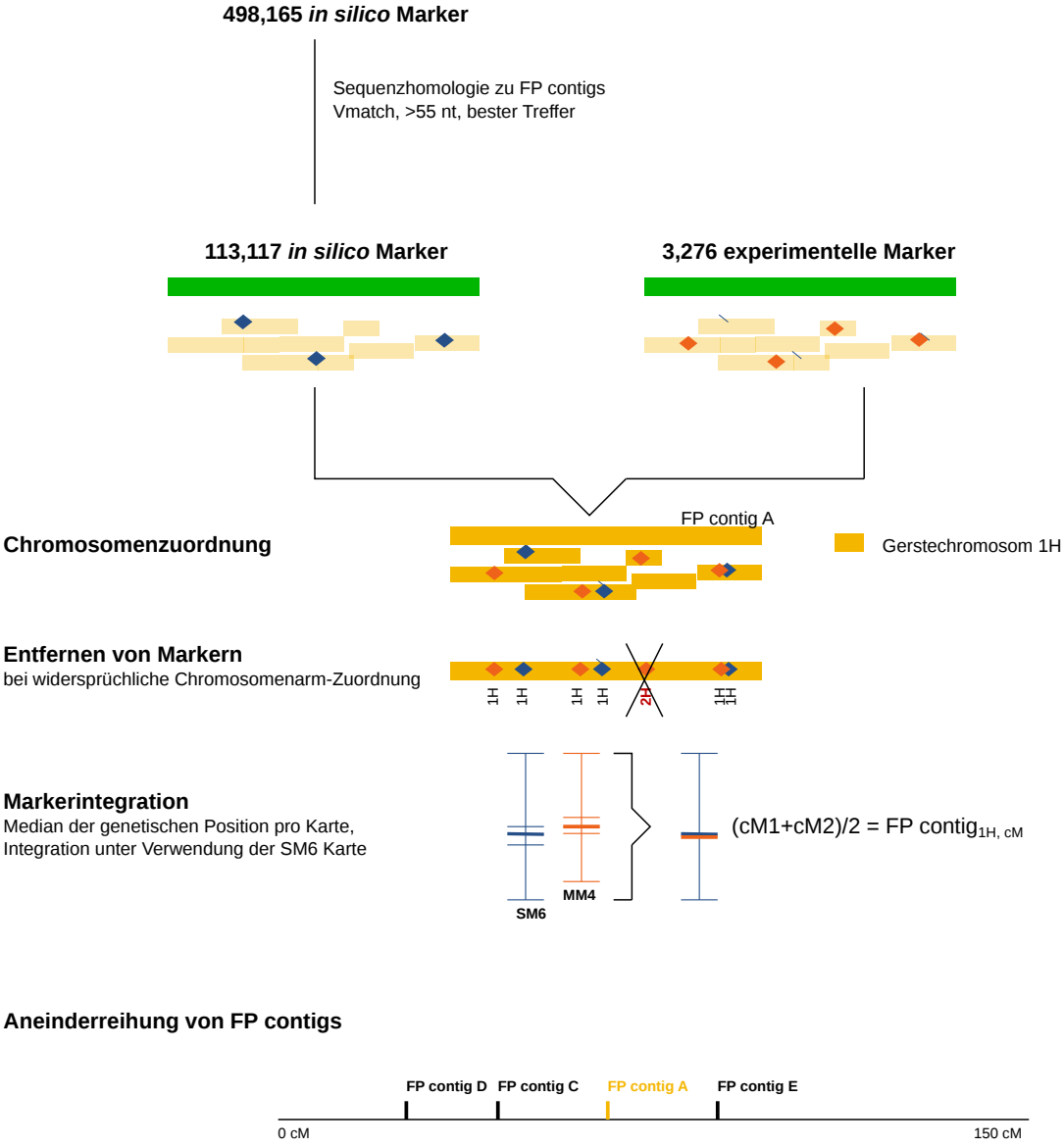

Abbildung 2.1: **Markerkartenintegration in Gerste**.
In silico Marker und genetische Marker werden auf Sequenzen der physikalischen Contigs verankert. Die Chromosomenzuordnung des physikalischen Contigs erlaubt widersprüchliche Marker zu entfernen. Anschließend werden die Markerkarten zusammengefasst und physikalische Contigs linear angeordnet.

chromosom 6A wurde wie in Gerste durchgeführt: Im Gegensatz zu Gerste wurden allerdings ausschließlich *in silico* Marker genutzt und getrennt voneinander auf die Klone der physikalischen Karte übertragen. Für Treffer mit einer Mindestlänge von 55 bp wurde erneut Vmatch zur Suche nach dem

besten Treffer angewandt. Marker der Poland et al. [2012] Karte wurden in die Saintenac et al. [2013] Karte integriert, indem die genetischen Positionen mit dem Faktor 1.2 multipliziert wurden, der den Faktor aus dem Vergleich der kumulativen Längen der Markerkarten darstellt. FP contigs wurden gemäß der integrierten Karte angeordnet und physikalische Längen wurden aufsummiert, um das 6A-Pseudochromosom zu bestimmen.
Die Gerste wurde genutzt, um weitere FP contigs in Weizenchromosom 6A zu verankern. Zunächst wurden 6A FP contigs mit den Gersten NGS contigs aus POPSEQ (Mascher et al. [2013]) verglichen. Eine Mindestsequenzidentität von 87%, Mindestlänge von 300 bp und der beste bidirektionale Treffer zu einem Gersten NGS contig war Voraussetzung für eine Verankerung. 6A FP contigs wurden gegen 15,719 verankerte Gerstengene verglichen (IBSC [2012]). Eine Sequenzlänge von 30 Aminosäuren und Mindestsequenzidentität von 75% Sequenzidentität musste erfüllt sein und ein bester bidirektionaler Treffer zu einem Gerstengen musste vorliegen.

2.5 Klonassemblierung und Klondekonvolution in Weizen

59 Klonpools mit jeweils 384 Klonen, in Summe 7,064 Klone, stellten 45 Gb Rohdaten bereit. Sequenzen wurden mit Hilfe der Programme Bowtie (Langmead and Salzberg [2012], Standardparameter) von Kontaminationen (*E.coli*, NCBI Reference Sequence: NC_013361.1) und Vektorsequenzen (`http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/`) befreit. Verbliebene Sequenzen wurden mit Hilfe von SOAPdenovo (Luo et al. [2012]) unter Verwendung unterschiedlicher k-mer Größen (k=25, 40, 55, 70, 85) assembliert. Die weiteren Analysen basierten auf der Assemblierung mit einer k-mer Größe von 85. Dieses Assembly stellt die größte Sequenzmenge bereit (Sequenzen mit $\geqslant$200 bp; k-mer 25 (49 Mb), 40 (170 Mb), 55 (308 Mb), 70 (478 Mb), 85 (550 Mb)). Der längste erstellte Contig wies eine Länge von 41 kb auf, die durchschnittliche Contig-Länge betrug 531 bp und die mittlere (Median) Länge wies 624 bp auf. Die 550 Mb dieses Assemblierungsdatensatzes wurden anschließend zur Klondekonvolution genutzt. Dieser Vorgang wird in Kapitel 3.3.1 beschrieben. Im Zuge der Klondekonvolution wurden assemblierte Contigs bei Sequenzhomologietreffer durch die 59 Klonpools jeweils getrennt maskiert. Die Maskierung wurde mit Vmatch (`www.vmatch.de`, Option: -dbmaskmatch, l=100 bp) durchgeführt. In Folge wurden jene drei Klonpools mit größter Abdeckung auf einer bestimmten

Sequenz betrachtet und einem Klon zugeordnet, wenn die drei Pools jeweils aus der x-, y- und z-Achse stammten. Der Schnittpunkt aus den drei Achsen erlaubte eine Dekonvolution der Sequenz zu einem bestimmten Klon (siehe Tabelle 3.15).

2.6 Analysen zur Bestimmung der Genexpression in Weizen

Die Genomassemblierung von Weizen aus Brenchley et al. [2012] wurde als Referenzsequenz genutzt, um mit Hilfe von RNA-seq Daten Genmodelle in Weizen zu annotieren und anschließend deren Expressionswerte zu bestimmen (Kugler et al. [2013]). Die Zuweisung der RNA-seq Daten auf die Referenzsequenz erfolgte mit Hilfe von den Programmen Tophat und Cufflinks (Trapnell et al. [2012]). Es wurden 233,780 potentielle Gene bestimmt. Differentiell regulierte Gene wurden durch Verwendung der Cuffdiff-Methode (Trapnell et al. [2013]) bestimmt und zwar durch paarweise Vergleiche jeweils zwischen Pilzbefall und Negativkontrolle (Wasserbehandlung), getrennt in beide Zeitpunkte 30 und 50 Stunden nach Befall. Vmatch wurde genutzt, um diese potentiellen Gene zu Gerstengenen zuzuweisen und dadurch proteinkodierende Gene zu bestimmen. Eine Mindesttrefferlänge von 100 Nukleotiden sowie Mindestidentität von 75% wurde vorausgesetzt. Von 233,780 Genkandidaten wurden zwischen 15,360 (NIL1) und 15,797 (NIL2) Gerstengene über den besten bidirektionalen Treffer verbunden. Weizengenkandidaten mit einem Gerstenortholog wurden genutzt und die Expressionswerte über alle Bedingungen zur Erstellung eines Koexpressionsnetzwerkes verwendet, wenn zumindest eine Variationskoeffizient von 1 über alle Bedingungen überschritten wurde. Für das Genexpressionsnetzwerk wurde WGCNA (Langfelder and Horvath [2008] beta=4, Pearson-Korrelationskoeffizient) verwendet. WGCNA erlaubte es, Gene in Module zu gruppieren. Gene eines Moduls weisen eine ähnliche Expression über alle Bedingungen auf und wurden durch die "topological overlay matrix" (cutreeDynamic-Methode; deepSplit=2, Mindestmodulgröße von 40) in Module gruppiert. Anschließend wurden Genfamilien mit signifikant differentiell regulierten Genen ausgewählt und aus den 233,780 Genen Kandidaten extrahiert, die Sequenzhomologie zu den Sequenzmotiven wie WRKY, UGT, NBS-LRR und Glukanasen aufwiesen (vgl. Kugler et al. [2013]). Zur Erstellung des Sequenzalignments wurde ClustalW (Thompson et al. [2002]) genutzt und iTOL (Letunic and Bork [2011]) zur Darstellung des Dendro-

gramms genutzt.

2.7 Chromosomenarm-Zuordnung

Die Chromosomenarm-Zuordnungs-Methode (*CarmA*) bestimmt den wahrscheinlichsten Chromosomenarm für eine genomische Sequenz. Die Methode wurde zusammen mit Dr. Heidrun Gundlach entwickelt, durch Verena Prade erweitert und auf mehrere Pflanzen angewandt (Prade [2013]). Die Methode kann für jedes beliebige Genom angewandt werden, wenn Chromosomenarm sortierte Sequenzen vorliegen. In dieser Arbeit wurde sie exklusiv für Gerste genutzt und Chromosomenarm sortierte 454-Sequenzen aus Mayer et al. [2011] genutzt. Anschließend wurde Vmatch (`www.vmatch.de`) verwendet, um Sequenzen ohne Chromosomenzuordnung durch Sequenzen der 454-Sequenzierung der einzelnen Chromosomenarme zu maskieren. In Weizen wurden in Chromosomenarme sortierte NGS contigs des IWGSC [2014] genutzt und mit den assemblierten Klonsequenzen aus den Weizenchromosomen 1D, 4D und 6D verglichen. Damit wurden BAC-Klone, die aus einem einheitlichen Datensatz stammten, in die drei Chromosomen getrennt. Für Gerste und Weizen wurden 100 bp Mindesttrefferlänge und Maximalfehler von zwei verwendet und eine Zuordnung durchgeführt, wenn die beste Chromosomenzuordnung eine um zumindest 20% höhere Abdeckung auf der zu untersuchenden Sequenz erreichte, als der zweitbeste Treffer.

2.8 Programme zur Datendarstellung

2.8.1 *RNASeqExpressionBrowser*

Zur Analyse der Expressionsdaten von Weizengenen bei Pilzbefall wurde ein Webtool entwickelt (Nussbaumer et al. [2014a], Abbildung 2.2). Basierend auf einem Pythonskript werden Expressions-, Sequenz- und funktionelle Daten in eine MySQL-Datenbank geladen. Die Expressionswerte werden in Form einer Textdatei importiert. Zeilen stellen jeweils Gene dar und Spalten definieren die integrierten Bedingungen. Zusätzlich kann Domain-, Sequenz- und beschreibender Text angeführt werden. Der *RNASeqExpressionBrowser* erlaubt außerdem eine Projektbeschreibung als HTML-Code zu integrieren und ermöglicht Verweise zu externen Datenquellen und anderen Studien (*u. a. MIPS* PlantsDB (Nussbaumer et al. [2013]). Der *RNASeqExpressionBrowser* wurde auch in Kugler et al. [In Vorbereitung] und in Dey et al.

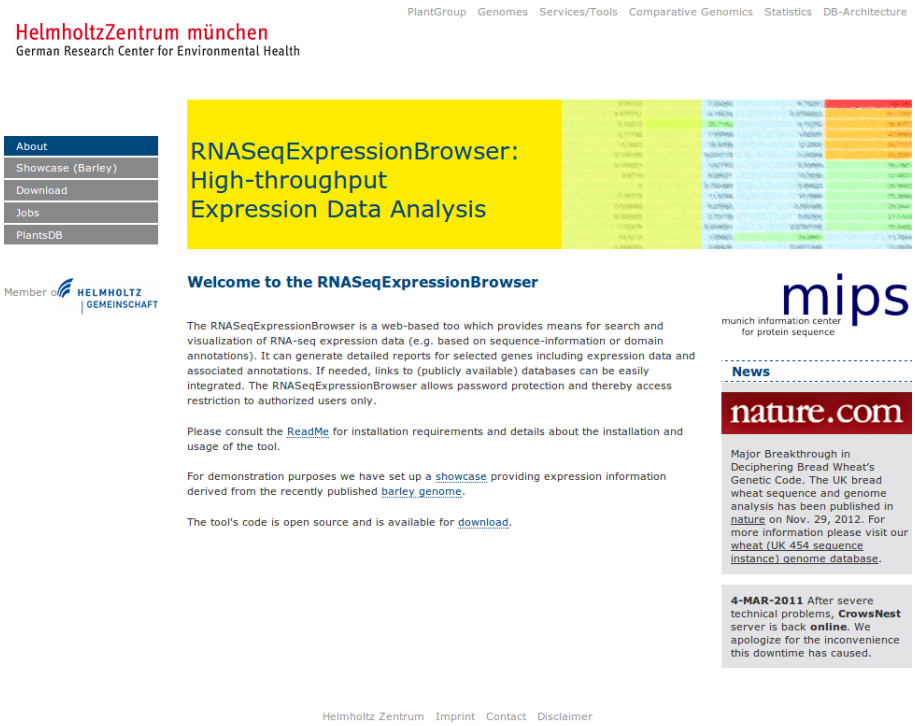

Abbildung 2.2: ***RNASeqExpressionBrowser***. Startseite des RNASeqExpressionBrowser `http://mips.helmholtz-muenchen.de/plant/RNASeqExpressionBrowser/`. Zur Illustration wurden Expressionsdaten aus Gerste dargestellt. Das Programm kann außerdem kopiert und lokal installiert werden. Zur Illustration sind Expresssionsdaten aus Gerste dargestellt.

[2014] eingesetzt.

2.8.2 GBrowse und FP contig-Darstellungen in Gerste

Für die Darstellung von Klonen und den darauf verankerten Sequenzen wurden Visualisierungsprogramme entwickelt, um die Fülle an FP contig-assoziierter Information zusammenzufassen, sowie den Austausch mit biologischen Kollaborationspartnern zu ermöglichen. Während der Verbesserung der Genom- und Klonassemblierungen wurden sie eingesetzt und halfen bei der Suche nach schimärischen FP contigs. Mit GBrowse wurden alle FP contigs (mit oder ohne genetischer Position) als eigene Entität dargestellt. CrowsNest (Nussbaumer et al. [2013]) wurde um eine vereinfachte FP contig-Darstellung erweitert, realisiert in der Programmiersprache Perl. Beide Visualisierungen dienten zum internen Austausch der jeweiligen Zwischenversionen der FP contig-Verankerungen und verankerten Ressourcen.

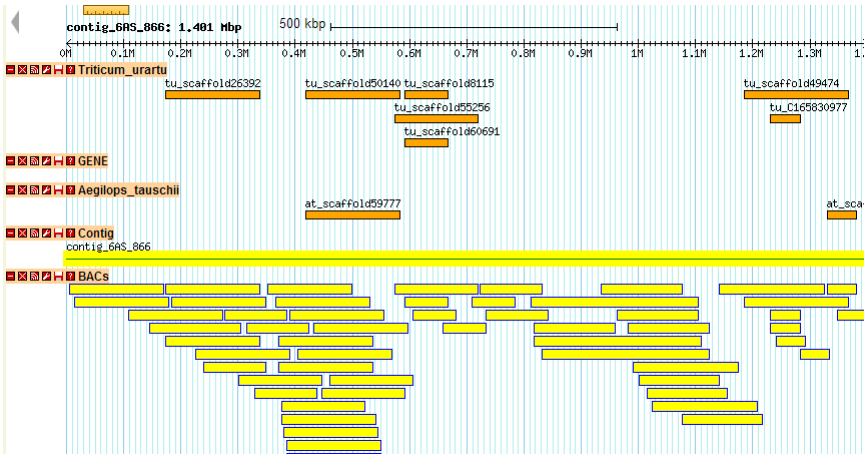

Abbildung 2.3: **GBrowse-Darstellung für Weizen**.
Darstellung des mit LTC assemblierten FP contigs 866 auf Weizenchromosom 6AS. Darstellung verankerter Gene (Spur: GENE) des *IWGSC* sowie der FP contig-assoziierten NGS contigs der Vorläufergenome (*Triticum urartu*, *Aegilops tauschii*). Der mit LTC erstellte FP contig contig_6AS_866 weist eine Länge von 1.401 Mb auf.

2.8.3 GBrowse und FP contig Darstellung in Weizen 6A

Für die Visualisierung der mit LTC erstellten FP contigs aus Weizenchromosom 6A wurden FP contigs samt assoziierter Sequenzen in GBrowse integriert (`http://seacow.helmholtz-muenchen.de/cgi-bin/gb2/gbrowse/Wheat_PhysMap_6A`). Das Hauptelement der Visualisierung ist ein FP contig. Zusätzlich werden Klone dargestellt, sowie die darauf verankerten NGS contigs und Sequenzen aus den Vorläufergenomen (*Triticum urartu*, *Aegilops tauschii*) (siehe Abbildung 2.3).
Eine Sequenzhomologiesuche wurde implementiert und bei einer starken Sequenzhomologie zu einer Sequenz eines FP contigs wird der entsprechende Bereich in GBrowse angezeigt. Die Stringenz der Sequenzhomologie kann durch Auswahl des BLAST *e*-Values eingestellt werden.

Kapitel 3

Ergebnisse

3.1 Verankerung der physikalischen Karte des Gerstengenoms

Die Bestimmung der vollständigen Sequenz eines hoch-repetitiven Genoms ist eine besondere Herausforderung: Zum einen liefert die Assemblierung von NGS Einzelsequenzen nur kurze Konsensussequenzen, hohe Rechenkapazitäten für die Assemblierung sind notwendig, und die finanziellen Einschränkungen für die Sequenzierung erfordern sorgfältige strategische Überlegungen bezüglich der Qualität und Auswahl von Sequenziertechnologien und Sequenzier-Strategien.
Die zentrale Aufgabe dieser Arbeit war die Erstellung einer detaillierten Landkarte des Gerstengenoms mittels der Integration einer sehr großen Anzahl heterogener, bereits publizierter als auch für diese Studie zusätzlich erzeugter Daten. Ein erster Meilenstein in Richtung eines vollständigen Genoms wurde mit der Gründung des Internationalen Gerstengenom Sequenzierungskonsortiums (Schulte et al. [2009]) und mit der Erstellung von fünf BAC-Klonbibliotheken (Schulte et al. [2011]) erreicht, die eine wesentliche Grundlage für die in dieser Arbeit durchgeführten Analysen darstellt. Kapitel 3.1.2 dieser Arbeit beschreibt die Erstellung der FP contig-Karte, die durch Dr. Ruvini Ariyadasa am *IPK* Gatersleben durchgeführt wurde. Nachdem Klone zu FP contigs zusammengefasst wurden, sind Marker notwendig, um FP contigs anzuordnen und aneinanderzureihen (siehe Kapitel 2.2). Marker, die über Hybridisierung einem Klon zugewiesen werden, wurden als experimentelle Marker bezeichnet. Bei *in silico* Markern erfolgte eine Zuordnung zu einem Klon über Sequenzhomologie zu Sequenzen auf dem

FP contig. Das Vorgehen, wie einzelne heterogene Datensätze zur veranker-

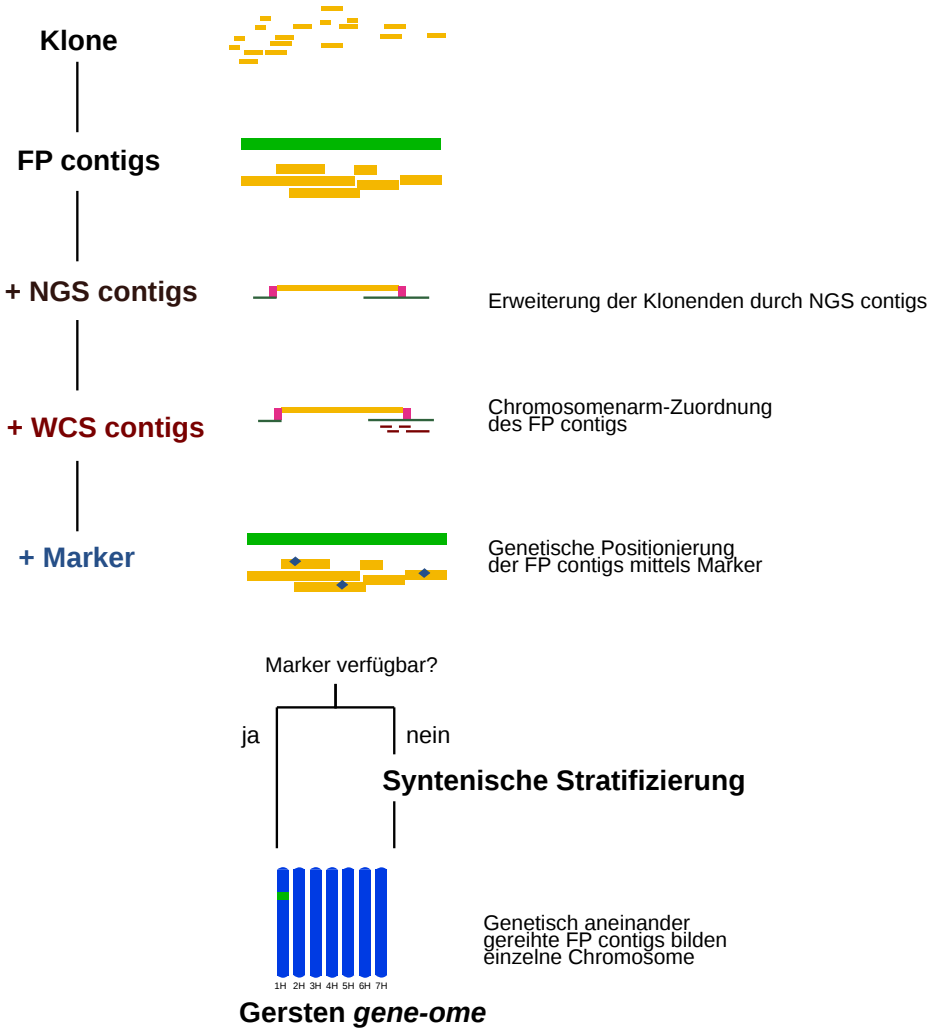

Abbildung 3.1: **Datenintegration.**
Schematische Darstellung der einzelnen Schritte der Datenintegration die zur verankerten physikalischen Karte von Gerste führten. WCS contigs = In Chromosomenarme sortierte contigs, Marker = *in silico* und experimentelle Markerkarten.

ten physikalischen Karte verbunden wurden, besteht aus mehreren Arbeitsschritten (siehe Abbildung 3.1): Im ersten Abschnitt wurden Klone zu FP contigs zusammengefasst. Für diese FP contigs liegen sequenzierte Klonenden und sequenzierte Klone vor und erlauben die Verankerung von NGS contigs aus drei Gerstenkultivaren über Sequenzhomologie. Die NGS contigs aus Gerstenkultivar “Morex” werden für die Genannotation genutzt, wodurch die physikalische Karte auch bereits mit Genen versehen wurde.

Auch für die Verwendung von BAC-Kibliotheken wurde der Gerstenkultivar Morex verwendet. Die Annotation mit Sequenzen der physikalischen Karte resultiert dann in einer ersten Anordnung und Assoziation von Genen in der Gerste. Der zweite Abschnitt beschreibt die Verankerung von FP contigs durch experimentelle und *in silico* Marker. FP contigs, ohne zugewiesener genetischer Position via Markerassoziation wurden über Synteniekriterien einer genomischen Position zugewiesen. Die Methodik wird als "syntenische Stratifizierung" bezeichnet und nutzte die kleineren und nahezu vollständig sequenzierten Referenzgenome Reis, *Sorghum* und *Brachypodium* (siehe Kapitel 3.1.7). Um die genetische Verankerung zu erleichtern, wurden FP contigs zunächst einem Chromosomenarm zugewiesen und anschließend wurden sie genetisch und genomisch verankert. Das Verfahren, das diese Zuordnung durch Berücksichtigung vom Chromosomenarm sortierten Sequenzen ermöglicht, wird als *CarmA* bezeichnet (Kapitel 3.1.5).

3.1.1 Datenquellen

Die Datensätze, die für die Erstellung des Gerste *gene-omes* genutzt wurden sind in ihrer Struktur äußerst heterogen und umfassen folgende Daten:

- FP contigs
- Klone
- Marker
- Hochdurchsatz-Sequenzdatensätze
- Genmodelle/Gensequenzen

Als FP contigs werden Einheiten verstanden, die abhängig von Genomgröße, Genomabdeckung und Gehalt an wiederholenden Basenpaarabfolgen eine bestimmte Anzahl überlappender Klone bündeln. Sieben unterschiedliche Klonbibliotheken mit in Summe 571,814 Klonen wurden für die FPC-Assemblierung verwendet. 6,200 zufällig ausgewählte Klone wurden sequenziert. Zum Projektstart wurden 3,158 Gen-tragende Klone (Madishetty et al. [2007]) und 517,202 sequenzierte Klonenden (*BES*) bereitgestellt, in weiterer Folge kamen 3,120 sequenzierte Klone hinzu (2,183 *UCR*-Klone (Lonardi et al. [2013]), 937 *IPK*-Klone (IBSC [2012])).
Marker stellen die nächste Datenquelle dar. Sie werden als experimentelle Marker bezeichnet und mit "MM" (***m**arker **m**ap*) abgekürzt oder als *in silico* Marker, wenn sie über Sequenzhomologie zugeordnet werden, abgekürzt durch "SM" (*in **s**ilico* **m**arker) (Tabelle 3.9). Ausgehend von veröffentlichten

Markerkarten wurden weitere Markerkarten (SM6, SM7-SM10, SM11) integriert. NGS Assemblies der Gerstenkultivare "Morex", "Barke" und "Bowman" mit unterschiedlicher Genomabdeckung ($50x$, $36x$ und $20x$) wurden erstellt und in die physikalische Karte des Gerstenkultivars "Morex" integriert. Für das Assembly von cv. "Morex" wurden Sequenzen aus Klonbibliotheken mit 500 bp und 2.5 kb Insertgrößen genutzt. Das Assembly wies eine kumulative Länge von 1.9 Gb auf, relativ zur erwarteten Genomgröße von 5.1 Gb (IBSC [2012]) und im Vergleich zu 1.8 Gb für cv. "Bowman" und 2.0 Gb für cv. "Barke". Die geringe Länge des Assemblies ist vor allem durch hoch-repetitive Bereiche bedingt. Repetitive Elemente liegen in Gerste und Weizen meist verschachtelt vor (Choulet et al. [2010]) und führen dazu, dass einzelne genomische Bereiche mit vielen repetitiven DNA Sequenzen kollabieren. Die Assemblierung des NGS Datensatzes aus Kultivar "Morex" wies 376,261 NGS contigs auf, die größer als 1 kb sind, bei einem *N50* von 1,425 bp (IBSC [2012]).
Annotierte Gene sind eine weitere Datenquelle der Verankerung. Gene wurden auf den NGS contigs annotiert unter Verwendung von 28,592 Gersten Volllängen-cDNAs und 834 Millionen RNA-seq Sequenzen (167 Gb), die aus acht Wachstumsstadien der Gerste stammen (IBSC [2012], Matsumoto et al. [2011]).
Klone sind direkt einem FP contig zugeordnet, während im Anschluss verankerte NGS contigs und Gene nur indirekt über Sequenzhomologie zu einem Klon auf die FPC-Karte übertragen werden. Die 6,200 sequenzierten Klone lagen nicht vollständig sequenziert vor, sondern in Form längerer, sequenzierter und assemblierter Contigs der Illumina- oder 454-Sequenzierung. In Summe stellten sie bis auf kleinere und nicht zu schließende Lücken, den gesamten Klon dar. Die Contigs können aber nicht in eine einzige, kontinuierliche Sequenz zusammengefasst werden (Steuernagel et al. [2009]). Die Assemblierung von 91 Klonen, bei einer $24x$ Genomabdeckung, wies beispielsweise maximal zehn Teilsequenzen auf.
Um Teilsequenzen in Klonen zu längeren Contigs zu assemblieren, ist die Verwendung von Sequenzbibliotheken mit größeren Abständen zwischen einander flankierenden Sequenzen ("paired-end reads" bzw. "mate-pair") notwendig. Alternativ werden Sequenziertechniken mit längeren Sequenzen (*z. B.* Pacific Biosciences) genutzt. Klone wurden in geringer Anzahl sequenziert, Klonenden hingegen für den Großteil der Klone bestimmt: 27,695 Klone besitzen ein sequenziertes Ende, in 175,831 Klone konnten beide Enden sequenziert werden. Für 2,276 Klone liegen mehr als zwei sequenzierte

Enden vor, resultierend aus Klonen, die in mehreren Bibliotheken berücksichtigt wurden (Bibliotheken der MOHI (HVVMRXALLHA) und MOKI (HVVMRX83KH)-Bibliothek). Die Erstellung der FP contig-Karte wurde von Dr. Ruvini Ariyadasa mit dem Programm FPC (Soderlund et al. [1997]) durchgeführt. Alternativ dazu wurde das Programm LTC (*Logical Topological Order*) (Frenkel et al. [2010]) genutzt. LTC wurde später auch genutzt, um für Klone in Gerste den minimal zusammengehörenden Pfad (MTP) auszuwählen (Ariyadasa et al. [2014]). Der Vergleich LTC/FPC zeigte, dass der Großteil der FP contigs in LTC Klone in einer linearen Struktur aneinanderreiht. Die lineare Struktur besagt, dass diese Klone auch in LTC zu einem FP contig geformt werden (Ariyadasa et al. [2014]).

3.1.2 Klonbibliotheken

Sieben Klonbibliotheken wurden für die Assemblierung in Gerste genutzt (Tabelle 3.1). 87% der Klone aller Bibliotheken wurden in FP contigs berücksichtigt. Selten vertreten sind Klone der Bibliotheken HVVMRXALLHC (10,324) und HVVMRXALLHB (38,238), während Klone der Bibliothek HVVMRXALLMA die Mehrheit bilden (161,905). Für eine bestimmte Klonbibliothek findet jeweils ein bestimmtes Restriktionsenzym Verwendung (Tabelle 3.2). Klone wurden entfernt, wenn Kontaminationen (*u. a.* bakterielle oder menschliche DNA) vorlagen oder eine zu geringe Anzahl an Restriktionsschnittstellen vorhanden waren.
Die größte Anzahl sequenzierter Klone wurde durch die mit *Hin*dIII behandelte HVVMRXALLH Bibliotheken bereitgestellt (4,253), gefolgt von der HVVMRX83KH Bibliothek (1,206), während Klone der restlichen Bibliotheken in geringen Mengen sequenziert wurden und in Summe 741 Klone beisteuern. Die Größe eines Klons wird in Konsensusbanden (*CB*) angegeben.
Die Einheit *CB* bezeichnet die kleinste Einheit der FP contig-Karte und umfasst eine bestimmte Anzahl an Fragmenten bei partiellem Verdau des Genoms durch ein oder mehrere Restriktionsenzyme.
Die Bestimmung der physikalischen Größe der *CB* Einheit erlaubt die Abschätzung der Länge eines FP contigs und damit der kumulativen Länge der FP contig-Karte. Dadurch wird überprüft, ob die Genomgröße erreicht wird und die FP contig-Karte jegliche genomische Bereiche umfasst. Tabelle 3.3 führt aufgetrennt auf einzelne Klonbibliotheken den Faktor zwischen der Gesamtlänge in bp und der *CB* Länge der jeweiligen Klonbibliotheken an:

Klonbibliothek	Abkürzung	FP contig Klone	verankerte Klone	% verankerte Klone	Restriktionsenzym
HVVMRXALLEA	MNEA	117,529	102,815	87	*EcoR*1
HVVMRXALLHB	MNHA	38,238	32,842	86	*Hin*dIII
HVVMRXALLHC	MNHB	10,324	9,069	88	*Hin*dIII
HVVMRXALLMA	MNMA	161,905	141,886	88	*Mbo*I
HVVMRXALLRA	MNRA	84,750	72,514	86	Mechanical sheared
HVVMRXALLHA	MOHI	39,378	32,876	83	*Hin*dIII
HVVMRX83KH	MOKI	65,078	59,109	91	*Hin*dIII
$\sum$	-	517,202	451,111	87	-

Tabelle 3.1: **Klonbibliotheken für die Assemblierung und Grad an genetischer Verankerung**.
517,202 Klone wurden für die Assemblierung von Klonen zu FP contigs genutzt und 87% der Klone konnten in dieser Studie schließlich genetisch verankert werden.

Restriktionsenzym	Sequenzmotiv
*Eco*RI	G-AATTC
*Mbo*I	GATC
*Hin*dIII	A-AGCTT

Tabelle 3.2: **In Gerste verwendete Sequenzmotive zur Erstellung der Klonbibliotheken**. *Hin*dIII wurde in vier Klonbibliotheken genutzt, *Eco*R1 in einer Klonbibliothek und *Mbo*1 ebenfalls in einer. Für eine Klonbibliothek wurde mechanisches Scheren genutzt.

Klonbibliothek	CB Länge	Klonanzahl	kb pro CB	bp/CB
HVVMRX83KH	95	1,206	122	1,287
HVVMRXALLE	104	309	142	1,366
HVVMRXALLH	100	4,253	112	1,122
HVVMRXALLM	118	256	170	1,445
HVVMRXALLR	107	176	131	1,226
$\sum$	104	6,200	131	1,255

Tabelle 3.3: **Bestimmung der Klongrößen**. Der Quotient aus der Gesamtlänge (in bp) zum *CB* Median sowie *CB* und kb Längen pro BAC-Bibliothek.

Für Klone der HVVMRXALLMA Bibliothek entspricht 1 *CB* 1,445 Basenpaaren, für Klone der HVVMRXALLH Bibliotheken 1 *CB* nur 1,122 bp. Über alle Klone gemittelt entspricht 1 *CB* 1,255 bp. Die FP contig-Karte erreicht somit eine kumulative Länge von 4.997 Mb. Für jede der sieben Bibliotheken wurde die Länge an *CB* Einheiten bestimmt, die von zumindest einem Klon einer bestimmten Bibliothek abgedeckt wird. Der Anteil der physikalischen Karte pro Klonbibliotheken kann daraus abgeleitet werden. Die Bibliothek HVVMRXALLMA sticht hervor: Alleine 4.4 Gb der 5.0 Gb werden mit dieser Bibliothek exklusiv verankert (Tabelle 3.4). 2,752 Mb werden durch HVVMRXALLHA Klone abgedeckt und 4,350 Mb mit der Bibliothek HVVMRXALLMA. Die Verbindung unterschiedlicher Bibliotheken ist notwendig, um alle genomischen Bereiche weitestgehend abzudecken. Außerdem erhöht die Anzahl an Klonen die Signifikanz von Klonüberlappungen, als Grundlage für einen robusten minimalen überspannenden Pfad (MTP und führt dazu, dass mehrere Klone zu einem FP contig zusammgenfasst werden.

Klonbibliothek	Kumulative Klonbibliothek Länge
HVVMRXALLMA	4,350
HVVMRXALL83KH	2,196
HVVMRXALLHC	1,078
HVVMRXALLHB	2,712
HVVMRXALLHA	2,752
HVVMRXALLRA	3,973
HVVMRXALLEA	3,982
$\sum$	4,997

Tabelle 3.4: **Anteil der Klonbibliotheken an der physikalischen Karte in Gerste**. Kumulative Länge einzelner Klonbibliotheken; nur einfache Genomabdeckung wird gezählt. Angaben in Mb.

3.1.3 Assemblierung von Klonen zu FP contigs

Die Assemblierung der Klone zu FP contigs wurde am *IPK* Gatersleben durchgeführt. Mit Hilfe des Programms FPC (Soderlund et al. [1997]) wurden sieben Klonbibliotheken mit 600,000 Klonen berücksichtigt und auf 571,814 Klone reduziert. Diese Klonanzahl repräsentiert eine annähernd 14-fache Genomabdeckung. Klone wurden entfernt, wenn nach Verdau durch die Restriktionsenzyme die Fragmentanzahl <30 oder >250 betrug. Eine zu kleine Fragmentanzahl würde im Zuge der Assemblierung nur zu gering signifikanten Klonüberlappungen führen und kann zu schimärischen FP contigs führen.
Die Klonbibliotheken weisen eine stark unterschiedliche Klonanzahl auf. Die Anzahl an Klonen, die für die Assemblierung pro Bibliothek schließlich berücksichtigt wurde, reicht von 10,324 Klonen für die Klonbibliothek HVVMRXALLHC bis hin zu 161,905 Klonen für die Bibliothek HVVMRXALLEA (siehe Tabelle 3.1). Die Klone, die in die FP contig-Assemblierung einflossen wurden im Laufe mehrerer Iterationen mit absteigendem Sulston Score (e^{-90} auf e^{-45}) assembliert. Die Anzahl an FP contigs konnte von 18,570 auf 9,436 reduziert werden. Durch die verschiedenen Iterationsschritte wurde die Anzahl an Klonen, die nicht zu FP contigs zusammengefasst (sogenannte "singletons") konnte von 89,849 auf 53,805 reduziert. Genetische Marker erlaubten eine manuelle Überprüfung von FP contigs. 171 FP contigs konnten durch widersprüchliche genetische Markerzuordnungen bestimmt und neu assembliert werden. Andererseits wurden unter Zuhilfenahme von experimentellen als auch *in silico* Markerzuordnungen 130 FP

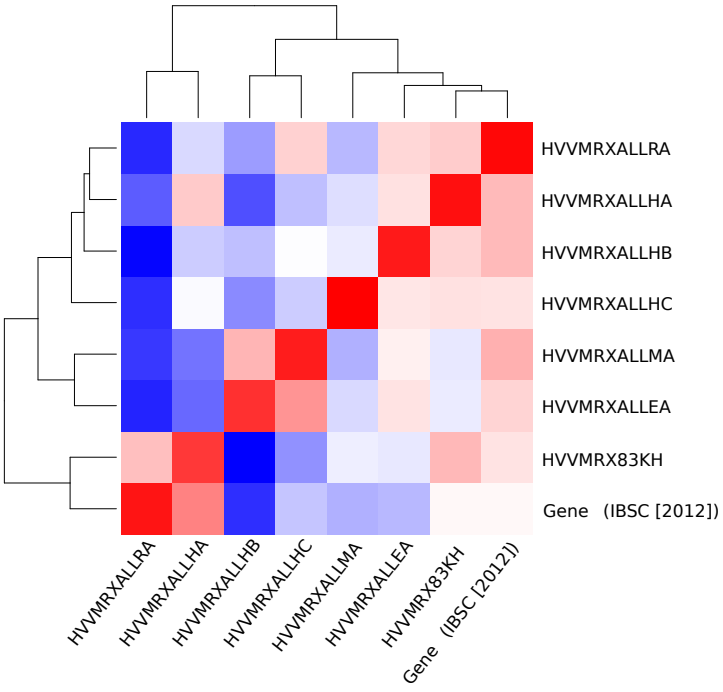

Abbildung 3.2: **Korrelationen zwischen Klonbibliotheken**.
Korrelationskoeffizient nach Bravais-Pearson zwischen einzelnen Klonbibliotheken und Gendichte bezüglich der Klonanzahl pro 10 Mb-Bereich. Rote Blöcke weisen auf eine starke Korrelation hin, blaue Blöcke auf eine schwache negative Korrelation. Der Wertebereich reicht von -0.14 (HVVMRXALLEA vs. Gene) bis 1 (*u. a.* HVVRXALLEA vs HVVRXALLEA).

contigs zusammengefasst. In dieser Arbeit wurden Analysen beigesteuert, um Dr. Ruvini Ariyadasa bei diesen Arbeitsschritten zu unterstützen. Die finale FP contig-Karte umfasst 9,265 FP contigs mit einer durchschnittliche FP contig-Länge von 538 kb. Die geschätzte Genomabdeckung wird von 5.01 Gb auf 4.99 Gb reduziert, die durchschnittliche FP contig-Länge von 515 auf 521 kb erhöht. In Ariyadasa et al. [2014] wurde die Verteilung von Klonen der einzelnen Klonbibliotheken entlang der physikalischen Karte untersucht: Die Anzahl an Klonen pro Mb wurde durch die erwartete Anzahl geteilt. Sechs der sieben Bibliotheken wiesen einen Faktor von eins über das gesamte Genom auf. Es bedeutet, dass die gemessene Klonanzahl pro Mb-Bereich der zu erwartenden Anzahl entspricht, während die besonders genreiche Bibliothek (HVVMRX83KH) eine Unterrepräsentierung von Klonen im zentralen Bereich des Chromosoms, dem Sitz des Zentromers, aufweist. Eine deutliche Überrepräsentierung dieser Bibliothek liegt in den distalen chromosomalen Bereichen und somit vergleichsweise genreichen Re-

gionen vor. In vielen Getreidengenomen weisen nämlich die distalen Bereiche des Chromosoms besonders hohe Gendichten auf (*u. a.* The International Brachypodium Initiative [2010], IRGSP [2005]). Es bewirkt, dass die übrigen sechs Bibliotheken dadurch in den distalen Bereichen unterrepräsentiert sind. In den Zentromer-nahen Bereichen hat die Unterrepräsentierung von HVVMRX83KH hingegen keinen großen Einfluss.

3.1.4 Integrationen der einzelnen Datenressourcen

Die FPC-Karte enthält die Information, welche Klone zu einem FP contig gehören und führt den Bereich in Konsensusbanden (*CB*) an, den ein Klon auf dem FP contig einnimmt. Die Orientierung des Klons im Bezug auf seine Klonierungsenden ist hingegen unbekannt. Um den Nutzen der FP contig-Karte maximal auszuschöpfen, reicht eine physikalische Karte ohne assoziierte Sequenzen aber nicht aus. Bisher wurden in Gerste nur 10% des Genoms sequenziert. Klonenden liegen für den Großteil der Klone vor, ihre beiden Endsequenzen weisen etwa 1,400 bp im Vergleich zu den Klonlängen von 100-200 kb auf. Die NGS contigs andererseits beschränken sich bei hochrepetitiven Genomen meist auf die wenig-repetitiven Bereiche und erreichen nur mittlere Längen von 1.5 kb je Contig. Die Klonenden sind gleichmäßig über das Genom verteilt und enthalten - proportional zum Grad der Genomrepetivität - mit hoher Wahrscheinlichkeit hoch-repetitive Sequenzen. Anschließend wurden NGS contigs der drei Gerstenkultivare “Morex”, “Barke” und “Bowman” auf die physikalische Karte übertragen (IBSC [2012], siehe Kapitel 2.3). In Summe konnten 5,798 (94%) sequenzierte Klone, 338,368 (59%) der Klonenden sowie 179,993 (87%) Klone mit sequenzierten Klonenden einem FP contig zugewiesen werden.

3.1.5 Chromosomenarm-Zuordnung (*CarmA*)

Im vorherigen Kapitel wurden den einzelnen FP contigs Sequenzinformation zugewiesen. Bevor FP contigs durch Marker eine genetische Position erhalten, wird eine Chromosomenarm-Zuordnung für den FP contig mit allen verankerten Sequenzen bestimmt. Die Zuordnung eines FP contigs zu einem Chromosomenarm ist aus drei Gründen notwendig:
Zunächst wird ein FP contig einem bestimmen Chromosomenarm zugeordnet, wodurch die genetische Verankerung des FP contigs vereinfacht wird, weil falsche Markerzuordnungen entfernt werden, die auf andere Chromosomen hinweisen. Andererseits werden Klone ersichtlich, die fälschlicherweise

einem FP contig zugeordnet wurden, ersichtlich durch eine widersprüchliche Chromosomenarm-Zuordnung zwischen Klon und zugehörigem FP contig. Schließlich können verankerte Gene entfernt werden, die fälschlicherweise einem FP contig zugewiesen wurden. Durch genetische Marker können diese Gene allerdings an die korrekte Position verankert werden.
Das dafür entwickelte Verfahren wird als *CarmA* (*Chromosome arm assignment*) bezeichnet. *CarmA* wurde zusammen mit Dr. Heidrun Gundlach und Verena Prade (Prade [2013]) entwickelt und wird in Kapitel 2.7 beschrieben. Ausgehend von in Chromosomenarme sortierten Einzelsequenzen wird die zu untersuchende und bisher nicht genetisch verankerte Sequenz durch alle Chromosomenarme getrennt maskiert. Anschließend wird die absolute Abdeckung (in bp) pro Chromosomenarm bestimmt. Erreicht ein Chromosomenarm die höchste Abdeckung und liegt ein zuvor definierter Faktor (*z. B.* 2) deutlich höher als das Verhältnis des besten zum zweitbesten Treffer, erfolgt eine Zuordnung zu diesem Chromosom.
Eine Erhöhung des Faktors des besten zum zweitbesten Treffer dient dazu, falsche Treffer auszuschließen, beispielsweise wenn eine (moderat) repetitive Sequenz einem Chromosomenarm zugeordnet wird und die Sequenz in allen Chromosomen eine hohe Abdeckung erreicht (IBSC [2012]).
In Gerste wurden in Chromosomenarme sortierte 454-Sequenzen bei einfacher Genomabdeckung bereitgestellt. *CarmA*-Bestimmungen wurden für FP contigs, NGS contigs und Gene durchgeführt. Nur Sequenzen mit eindeutiger Chromosomenarm-Zuordnung wurden für die Verankerung genutzt. Abbildung 3.3 illustriert das Prinzip von *CarmA*: Eine beliebige Sequenz wird durch alle Chromosomenarme getrennt maskiert. Anschließend werden die Chromosomenarme mit höchster Trefferanzahl auf der Sequenz berücksichtigt. Im konkreten Beispiel sind es “1H’ sowie “2H”. Der Faktor des besten zum zweitbesten Treffer ist im vorliegenden Beispiel $\frac{2H}{1H} = \frac{100}{18}$. Wird ein Faktor von >2 vorausgesetzt, wird die Sequenz “2H” zugewiesen. In hoch-repetitiven Regionen mit Treffern zu mehreren Chromosomenarmen wird eine Zuordnung erschwert. *CarmA* wurde in Prade [2013] auf einzelne Gräserarten angewandt, unter anderem auch für *Brachypodium*, *Sorghum* und Mais (Prade [2013]). Die Spezifität der *CarmA*-Zuordnungen wiesen für diese Genome Werte von > 95% für Sequenzen mit einer Länge von mindestens 500 bp auf.

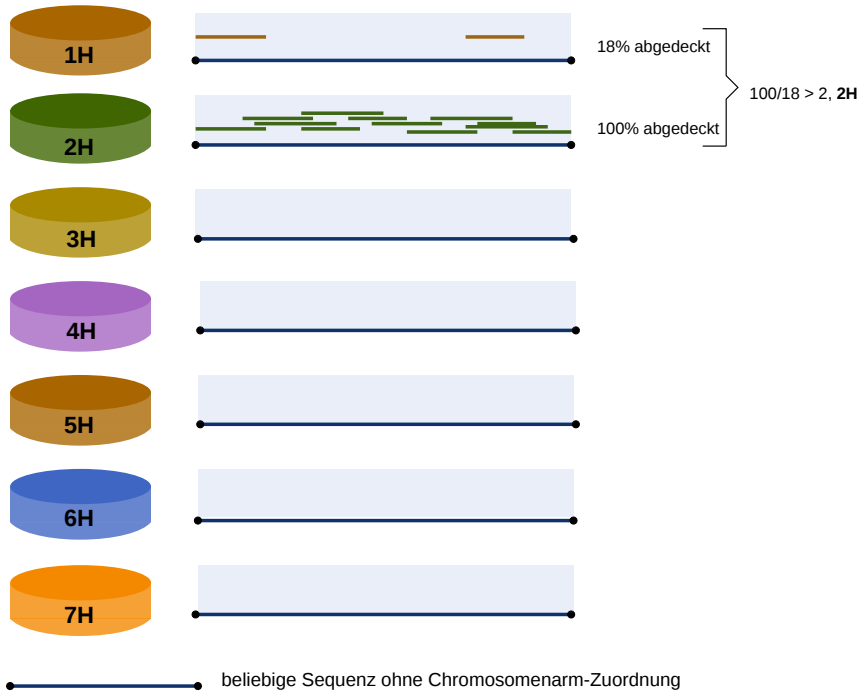

Abbildung 3.3: **Vorgehensweise von *CarmA***. Die Chromosomenarm sortierten Sequenzen aus Mayer et al. [2011] werden über Sequenzhomologie der zu untersuchenden Sequenz zugeordnet. Anschließend wird berechnet, welcher relative Anteil der Sequenz pro Chromosomenarm abgedeckt wird. Im konkreten Beispiel deckte Chromosom 2H die zu untersuchende Sequenz vollständig ab.

3.1.5.1 *CarmA*-Zuweisungen von FP contigs

Für die Chromosomenarm-Zuordnung der FP contigs wurden alle direkt (*NGS-*) bzw. direkt+indirekt (*NGS+*) verankerten Sequenzen genutzt. Direkt verankerte Sequenzen sind Klonenden und sequenzierte Klone, indirekt verankert sind NGS contigs der drei Kultivare “Morex”, “Barke” und “Bowman”. Um den Einfluss der Verankerung der NGS contigs auf die Arm-Zuordnung des zugrundeliegenden FP contigs abzuschätzen, wurde für FP contigs eine *CarmA* Bestimmung mit bzw. ohne verankerte NGS contigs durchgeführt. Die Fragestellung ist, ob die Berücksichtigung von NGS contigs die Qualität der Armzuordnungen erhöht und durch diese weiteren Sequenzen zusätzliche FP contigs einem Arm zugeordnet werden. In Gerste wird der umgekehrte Weg gewählt, NGS contigs mit einer FP contig Karte zu verbinden. In kleineren Genomen von <1 Gb wird die FP contig-Karte genutzt, um NGS contigs zusammenzufassen (Wang et al. [2014]). Die NGS contigs weisen bereits die Länge eines gesamten Chromosoms auf. Für Gerste

sind andererseits die Längen der FP contigs um ein Vielfaches größer als jene der NGS contigs. Die den FP contigs zugrundeliegenden Sequenzen sind aber deutlich repetitiver und werden größtenteils durch Klonenden definiert. Aus diesem Grund sollten *CarmA* zugeordnete NGS contigs die höchste Genauigkeit im Rahmen einer Chromosomenarm-Zuordnung aufweisen, da repetitive Bereiche in NGS contigs kollabieren und weniger repetitive ("low-copy regions") Bereiche verbleiben. Sequenzierte Klone liegen für 2,758 FP contigs vor, im Vergleich zu 9,265 FP contigs insgesamt. Für die restlichen FP contigs resultiert die Verankerung aus Homologie von Markern zu assoziierten Klonenden oder NGS contigs. Im Zuge der Erstellung des Gerste *gene-omes* (IBSC [2012]) und unter Zuhilfenahme aller FP contig assoziierten Sequenzen wurden 6,438 FP contigs einem Chromosom zugewiesen. Chromosom 1 in Gerste (1H) wurden 881 FP contigs zugeordnet. Die Anzahl an FP contigs, die einem anderen Chromosomenarm zugeordnet wurde, reicht von 370 (5HS) bis 589 (5HL). Tabelle 3.5 führt die Anzahl an FP contigs an, die mit Hilfe von direkt verankerten Sequenzen und ohne NGS contigs einem Arm zugeordnet wurden. Mit den 6,202 sequenzierten Klonen wurden lediglich 27.2% der FP contigs einem Chromosomenarm zugewiesen, mit Hilfe von Klonendsequenzen sind es hingegen 68.4% der FP contigs. Klonenden und sequenzierte Klone erlauben eine Chromosomenarm-Zuordnung von weiteren 51 FP contigs. Eine etwas höhere Anzahl an FP contigs wird chromosomal verankert, wenn auch NGS contigs berücksichtigt werden (6,438 im Vergleich zu 6,335). Tabelle 3.6 listet die Anzahl an in Chromosomenarme

Datenressource	Sequenzanzahl	FP contigs	% FP contigs[1]
sBAC	29,645	2,525	27.2%
BES	84,677	6,335	68.4%
sBAC+ BES	114,322	6,386	68.9%

Tabelle 3.5: **Chromosomenarm-Zuordnung für *BES* und *sBAC*.**
Anzahl an *BES*/*sBAC* mit einer Chromosomenarm-Zuordnung.
[1]bezüglich aller 9,265 FP contigs.

sortierte FP contigs, so wie sie schließlich unter Zuhilfenahme aller Sequenzen durch das Internationale Gestengenomprojekt im Rahmen dieser Arbeit bereitgestellt wurde (IBSC [2012]). Um den Einfluss auf die Verankerung von FP contigs durch Klonenden alleine zu untersuchen, wurde für *CarmA* lediglich Klonendsequenzen berücksichtigt (Tabelle 3.7). 342 (5HS) bis 553 FP contigs (5HL) wurden einem Chromosomenarm zugewiesen. Chromosom

Chromosomenarm	FP contigs insgesamt	Verankerte FP contigs
1H	881	560
2HS	428	323
2HL	503	392
3HS	378	270
3HL	516	398
4HS	439	251
4HL	518	345
5HS	370	257
5HL	589	431
6HS	409	298
6HL	486	346
7HS	508	377
7HL	412	308
$\sum$	6,438	4,556

Tabelle 3.6: **Chromosomenarm-Zuordnung für *BES* und *sBAC*.** Chromosomenarm sortierte FP contigs aus IBSC [2012] basierend auf verankerten *BES*, NGS contigs und *sBAC*.

1H, dem Chromosom für das keine Trennung in einzelne Chromosomenarme möglich war, wurden 927 FP contigs zugewiesen. Außerdem wurden die *CarmA*-Zuweisungen der FP contigs über Klonenden den publizierten Armzuweisungen der FP contigs aus IBSC [2012] gegenübergestellt. Der Vergleich zeigt, dass Klonenden alleine bereits ausreichen, um den Großteil der FP contig einem Chromosomenarm zuzuweisen, während weitere Sequenzen helfen, um die Chromosomenarm-Zuweisung zusätzlich zu bestätigen bzw. falsche Chromosomenarm-Zuweisungen zu lösen. Im Zuge der Bestimmung von schimärischen FP contigs wurde *CarmA* ebenfalls eingesetzt. Mit unterschiedlichen Stringenzkriterien (Mindestlänge eines Treffers, Verhältnis der Abdeckung zwischen besten zum zweitbesten Treffer) wurde eine *CarmA*-Zuordnung für Klone bestimmt.

3.1.5.2 *CarmA* für NGS contigs und Gerstengene

Für Gerste wurden drei wirtschaftlich wichtige Kultivare sequenziert und assembliert (IBSC [2012]) und im Rahmen dieser Arbeit mit Hilfe von *CarmA* einzelnen Chromosomenarmen zugewiesen (siehe Kapitel 2.7). Die Anzahl an

Chromosom	FP contigs	FP contigs verankert	Übereinstimmung[1]	Widerspruch[2]
1H	927	584	523	61
2HS	419	304	287	17
2HL	515	384	368	16
3HS	412	294	257	37
3HL	561	419	371	48
4HS	425	244	235	9
4HL	500	330	315	15
5HS	342	231	226	5
5HL	553	402	395	7
6HS	370	273	268	5
6HL	447	315	310	5
7HS	461	347	340	7
7HL	403	293	289	4
$\sum$	6335	4420	4184	248

Tabelle 3.7: **Chromosomenarm-Zuordnung von FP contigs rein auf Basis von Klonenden**.
In Chromosomenarme sortierte FP contigs unter Verwendung von ausschließlich *BES*.
[1]Gleiche Chromosomenarm-Zuweisung wie IBSC [2012].
[2]Widersprüchliche Chromosomenarm-Zuweisung zu IBSC [2012].

zugeordneten NGS contigs liegt für cv. “Morex” bei 307 Mb, für cv. “Bowman” bei 312 Mb und cv. “Barke” bei 268 Mb. NGS contigs aus Kultivar “Morex” wurden als Genomreferenz verwendet, um mit Hilfe von RNA-seq Daten (IBSC [2012]) und unter Zuhilfenahme von Gersten Volllängen-cDNA (Matsumoto et al. [2011]) Gene zu bestimmen. 26,159 HC (*high-confidence*) und 53,220 LC (*low-confidence*) Gene wurden annotiert und einem Chromosomenarm zugewiesen. Die Chromosomenarm sortierten Sequenzen erlauben es, die Gene einem Chromosomenarm zuzuweisen. Dadurch, und in Verbindung mit den rund 20 syntenischen Bereichen zu den kleineren Referenzgenomen erlaubt es bereits Aussagen, ob ein Gerstengen im erwarteten syntenischen Bereich liegt, oder ob es seine Lage beispielsweise durch Translokationen geändert wurde. Insgesamt erlaubt *CarmA* die Chromosomenarmzuordnung von 22,134 (84.6 %) HC-Genen sowie von 31,978 (60.1%) LC-Genen (Tabelle 3.8).

Chromosom	HC-Gene	LC-Gene	LC+HC Gene
1H	4,677	3,111	7,788
2HS	1,733	1,287	3,020
2HL	2,777	2,275	5,052
3HS	1,863	1,244	3,107
3HL	3,066	2,245	5,311
4HS	1,566	1,004	2,570
4HL	1,948	1,585	3,533
5HS	1,535	805	2,340
5HL	3,630	2,675	6,305
6HS	1,895	1,118	3,013
6HL	1,773	1,420	3,193
7HS	3,011	1,716	4,727
7HL	2,504	1,649	4,153
$\sum$	22,134	31,978	54,112

Tabelle 3.8: **Chromosomenarm-Zuordnung von Gerstengenen**. Die Zahlen führen alle annotierten Gene und wurden in HC (*high-confidence*) Gene und LC (*low-confidence* Gene) aufgetrennt.

3.1.6 Verankerung der FP contigs

Die FP contigs mit allen darauf zugewiesenen Sequenzen wurden über experimentelle und *in silico* Marker verankert und die Markerkarte mit der größten genetischen Auflösung als Referenzkarte genutzt (SM6, Comadran et al. [2012]). Die *in silico* Markerkarte basiert auf 3,973 polymorphen SNPs der Kreuzung Morex x Barke und wurde genutzt, um eine F_6 rekombinate Inzuchtlinien (RIL, *recombinant inbred lines*) Population bestehend aus 360 Individuen zu kartieren (Comadran et al. [2012]). Die Verankerungen der Marker aller anderen Karten wurden gegen die genetische Position der SM6-Karte verglichen, wenn diese Marker auf derselben Sequenz (FP contig, Gen, NGS contig) verankert wurden. Paarweise Vergleiche der genetischen Positionen einer beliebigen genetischen Karte gegen die Referenzkarte erlaubte die Bestimmung einer Funktion, die eine bestimmte genetische Position der einen Karte in eine genetische Position der Referenzkarte transformiert (siehe Kapitel 2.3). Eine mehrfache Wiederholung dieses Vorganges lieferte eine sogenannte integrierte genetische Karte, die Informationen aus einzelnen Karten in ein einheitliches Koordinatensystem abbildet. Für die Verankerung werden *in silico* und experimentelle Marker berücksichtigt. *In*

silico Marker bilden den Großteil der verwendeten Marker (99.3 %), experimentelle Karten umfassen 3,276 Marker. *GBS* Marker stellen die größte Sequenzmenge bereit (96.7%). Tabelle 3.9 führt die Anzahl an Marker für jede Markerkarte an. Eine besonders große Anzahl an experimentellen Mar-

Karte	Markeranzahl	Markertyp	Publikation
MM1	230	exp.	Chen et al. [2010]
MM2	297	exp.	Mayer et al. [2011]
MM3	1,736	exp.	Sato et al. [2009]
MM4	145	exp.	Stein et al. [2007]
MM5	596	exp.	Close et al. [2009]
MM6	342	exp.	Potokina et al. [2008]
$\sum$	3,276	exp.	
SM2	622	*in silico*	Mayer et al. [2011]
SM3	5,780	*in silico*	Sato et al. [2009]
SM4	1,052	*in silico*	Stein et al. [2007]
SM5	2,944	*in silico*	Moscou et al. [2011]
SM6	5,481	*in silico*	Comadran et al. [2012]
SM7	241,159	*in silico*	Poland et al. [2012]
SM8	34,396	*in silico*	Poland et al. [2012]
SM9	184,796	*in silico*	Poland et al. [2012]
SM10	21,384	*in silico*	Poland et al. [2012]
SM11	501	*in silico*	Moscou et al. [2011]
$\sum$	498,165	*in silico*	

Tabelle 3.9: **Markenkarten zur genetischen Verankerung von FP contigs**. Markeranzahl pro Markerkarte. Entnommen aus (IBSC [2012], 2012). exp. = experimentelle Marker. *in silico* = *in silico* Marker.

kern weist die Okayama Karte (SM3/MM3, Sato et al. [2009]) auf, die auf EST-Sequenzen (partielle Sequenzen aus cDNA) beruht, der SM5-Karte, die eine Konsensuskarte darstellt und zur Verankerung des GenomeZippers in Gerste genutzt wurde (Mayer et al. [2011]), und in der iSELECT Karte (SM6-Karte). Abbildung 3.4 illustriert, wie einzelne Markerkarten zu einer Karte verknüpft werden: Beginnend mit der Referenzkarte, werden weitere Karten durch gemeinsame Ankerpunkte zur Erstellung einer hochauflösenden Konsensuskarte für die Verankerung der FP contigs verwendet. Marker, die einem Transkript zugeordnet werden bieten die Möglichkeit, die Verankerung mit der genomischen Position der orthologen Gene zu vergleichen und

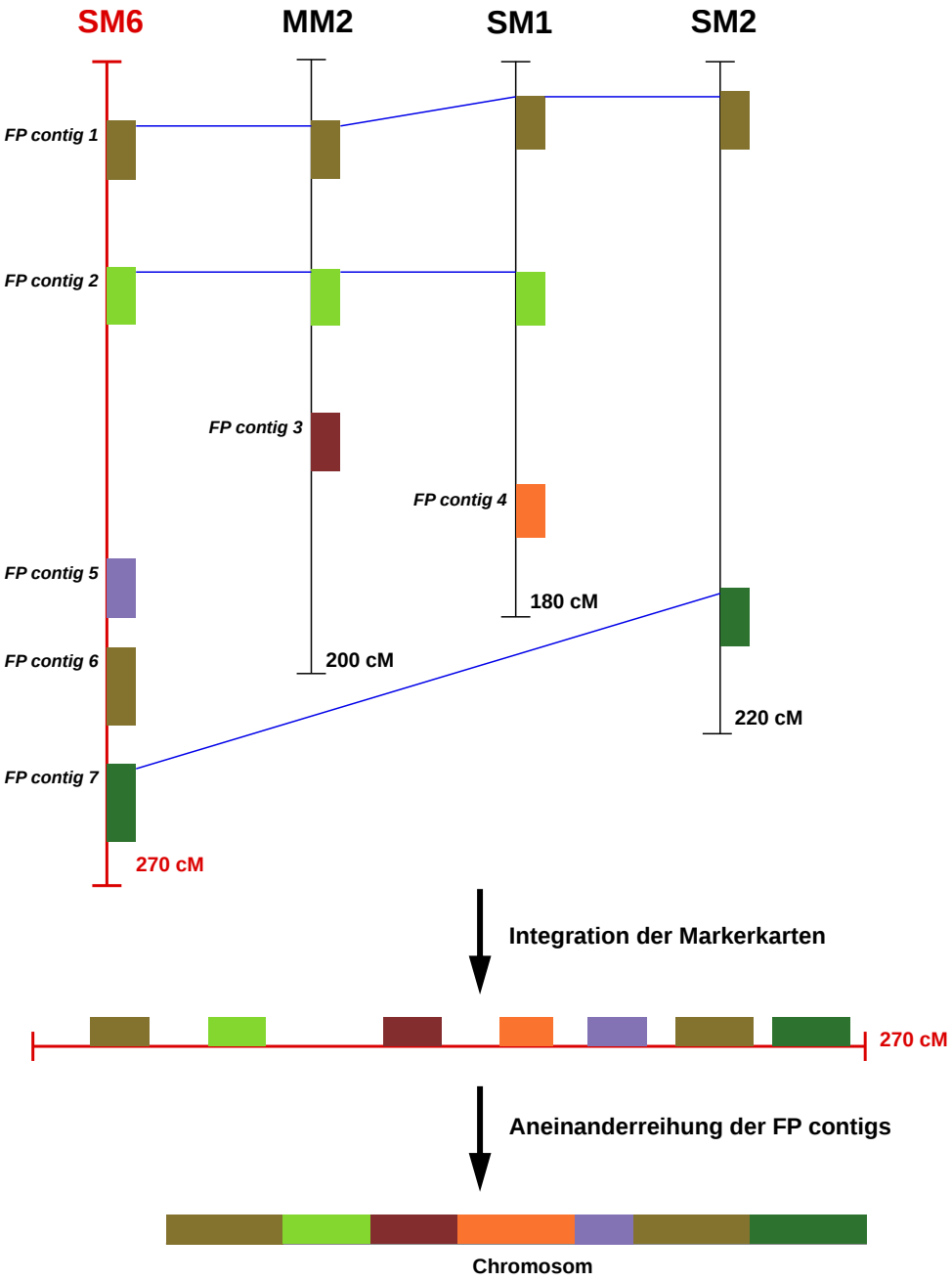

Abbildung 3.4: **Integration von Markenkarten und Verankerung von FP contigs**. Darstellung der Überführung mehrerer Markerkarten in eine Konsensuskarte. SM = *in silico* Marker, MM = experimentelle Marker. Die mit Rot versehene SM6-Karte stellt die Referenzkarte dar. Zunächst werden Marker einem FP contig zugeordnet. Anschließend die Markerkarten in eine hochauflösende integrierte Karte überführt und diese für die Aneinanderreihung de FP contigs genutzt.

damit eine grobe Einschätzung der Verankerungsqualität. Viele Markerkarten beruhen auf genischen Sequenzen. Die Anzahl an Markern, die einem Gen zugeordnet wurde, kann mittels Vergleich gegen die annotierten Gene des IBSC [2012] bestimmt werden (Tabelle 3.10). *GBS*-basierte Marker

Markerkarte	Genanzahl
SM2	128
SM3	1,766
SM4	250
SM5	1,327
SM6	3,405
SM7	3,323
SM8	1,189
SM9	2,791
SM10	1,068
SM11	392
$\sum$	8,809

Tabelle 3.10: **Markerkarten und Überlappung mit Gerstengenen**. Marker mit Sequenzhomologie zu einem der 26,159 annotierten Gerstengene. Die größte Genanzahl liegt in der Referenzkarte (SM6) vor.

(SM7-SM10) erlauben es, größtenteils nicht-genische Regionen zu verankern und treffen auf eine in Anbetracht der Markermenge relativ geringe Genanzahl. Die SM5-Karte erlaubt die Verankerung von 1,327 Gerstengenen. Die größte Genanzahl (3,405) wird mit der SM6 Karte erreicht. Trotz 241,159 Markern für die SM7 Karte liegen nur Treffer zu 3,323 Gerstengenen vor. Die Enzymkombination, die für die *GBS* Technologie genutzt wird, sorgt dafür, dass genflankierende Bereiche angereichert werden, nicht aber genische (Poland et al. [2012]). Es erklärt die geringe Überlappung mit genischen Sequenzen. In Summe besitzen 8,809 Gene eine eindeutige, genetische Position. Neben der Überlappung der Marker mit Genen und der Markerzahl ist die Länge der Markersequenzen ein entscheidender Faktor: Sind die Längen sehr kurz, wird eine Zuordnung zu einem Klon, Klonende oder NGS contig erschwert, während besonders lange Markersequenzen die eindeutige Zuordnung zu einer genomischen Position erlauben.

3.1.7 Syntenische Stratifizierung

Durch die Verankerung der FP contigs und die Überführung in eine gemeinsame Markerkarte wurden in dieser Arbeit 4,556 FP contigs mit einer summierten Länge von 3.9 Gb - etwa 78.2% der Gesamtlänge (4.99 Gb) aller FP contigs - einer genetischen Position zugewiesen. Neben diesen 4,556 FP contigs wurden 5,798 sequenzierte Klone und 338,368 Klonenden einer genetischen Position zugewiesen. Durch die syntenische Stratifizierung wird für 4,709 unverankerte FP contigs versucht, auch diesen eine genetische Position zuzuweisen. Wie in der Einleitung beschrieben, weisen Genome wie Gerste, Reis, *Brachypodium* und *Sorghum* eine sehr ähnliche Genreihenfolge aufgrund eines gemeinsamen Vorläufergenoms und dem Fehlen von weiteren Genomduplikationen auf. Im Gegensatz zu Mayer et al. [2011] wurden in dieser Arbeit mehrere und eine (SM6) besser auflösende, genetische Karte genutzt. Die höher auflösende genetische Referenzkarte ermöglichte eine feinere Eingrenzung der syntenischen Regionen und den bisher unverankerten FP contigs eine ungefähre Position in Gerste zuzuweisen, bei starker Sequenzhomologie zu einem bestimmten Bereich in *Brachypodium*, Reis oder *Sorghum*. Die syntenische Stratifizierung, die im Rahmen dieser Arbeit entwickelt wurde, setzt sich aus drei Schritten zusammen:

- Die Zuordnung von Gerstengenen zu spezifischen Chromosomenarmen.
- Die Bestimmung der Orthologen der Gerstengene in Modellorganismen.
- Interpolation der Abfolge von Gerstengenen über syntenische Stratifikation.

Eine Chromosomenarm-Zuordnung liegt für 54,112 Gerstengene (LC- und HC-Gene) vor (Tabelle 3.8). 22,134 davon sind HC (*high-confidence*) Gene. Mit einer Trefferlänge von 100 bp, bei einer Sequenzidentität von $>$ 75% werden Gerstengene einem *Brachypodium*, Reis oder *Sorghum* Gen zugeordnet. Der beste bidirektionale Treffer wird berücksichtigt und zum Sequenzvergleich BLASTp genutzt (Altschul et al. [1990]). Die Verankerung von 15,719 Gerstengene erlaubte die Bestimmung von syntenischen Bereichen zu den Referenzgenomen *Brachypodium*, Reis und *Sorghum* und damit die genetische Verankerung weiterer Gene über Syntenie. Weist ein Gen einen Treffer zu einem Gen aus *Brachypodium*/Reis/*Sorghum* auf, wird der dem Gen zugehörige syntenische Block bestimmt. Der syntenische Bereich erlaubt es, dem Gen eine Position zuzuweisen, wenn orthologe Gene in unmittelbarer Nähe genetisch in Gerste verankert wurden. Die Einschränkung

Genom	Klonenden	Gene	Transkripte
Brachypodium distachyon	22,087	9,560	11,262
Hordeum vulgare	11,320	6,692	6,692
Oryza sativa	18,349	9,359	12,628

Tabelle 3.11: **Verankerung der physikalischen Karte in Gerste rein auf Basis von Klonenden**. Klonenden wurden gegen Transkripte der Referenzgenome verglichen. Die Anzahl an Transkripten und Genen ist angeführt, für die Sequenzhomologie zu Klonenden gefunden wurde.

auf den syntenischen Block ist notwendig, weil nur positionell konservierte Gene verankert werden können. Anschließend wird der Median der vorgeschlagenen genetischen Position aus allen drei Referenzgenomen für die Verankerung genutzt; bei Widersprüchen, beispielsweise wenn die vorgeschlagene genetische Position im Widerspruch zur *CarmA*-Zuordnung ist, erfolgte keine Zuordnung. Dieser Vorgang führte dazu, dass neben den rund 15,719 positionierten Gerstengenen weitere 3,745 Gerstengene eine Verankerungsposition erhalten. Die Ergebnisse der syntenischen Stratifizierung ist unter `ftp://ftpmips.helmholtz-muenchen.de/plants/barley/public_data/anchoring/syn_strat/syntenic_stratification.TXT` hinterlegt.

3.1.8 Syntenische Verankerung von FP contigs

Mit Hilfe von *CarmA* wurde gezeigt, dass eine Chromosomenarm-Zuordnung, die ausschließlich auf Klonenden beruht, den Großteil der FP contigs einem Chromosom zugeordnet werden kann. *CarmA* zeigte auch, dass die Armzuordnung eines FP contig bereits durch Klonenden ermöglicht wird. Klonenden sind annähernd gleichverteilt über den FP contig und über das Genom und liegen in größeren Menge vor. Die verankerte physikalische Karte in Gerste erlaubt nun zu untersuchen, welcher Grad an Verankerung alleine aufgrund der Klonenden ohne Hilfe von Markersequenzen und mit Hilfe von Synteny zu sequenzierten Referenzgenomen möglich ist. Damit wird ermittelt, ob ein Syntenie vertrauendes Vorgehen hilft oder sogar ausreicht, um die Verankerung einer physikalischen Karte zu erreichen, für das weder Marker noch vollständige Chromosomen vorliegen. Durch die genetisch verankerten Gersten FP contigs kann der Erfolg und Verlässlichkeit dieses auf Syntenie basierenden Vergleichs überprüft werden.

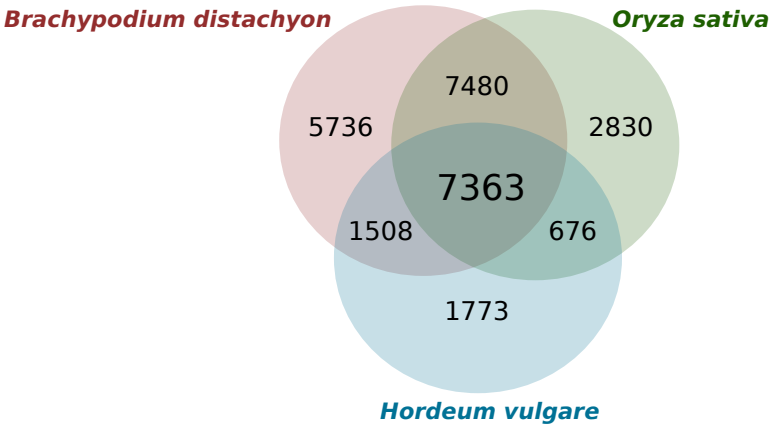

Abbildung 3.5: **Anzahl und Überlappung von Klonenden (*BES*) mit Sequenzhomologie zu *Brachypodium*, Reis und Gerste**.

Abbildung 3.5 zeigt den Anteil an Klonenden mit Sequenzhomologie zu *Brachypodium*, Reis sowie Gerste. Es wurden jeweils die kodierenden Sequenzen der drei Genome verwendet. 7,363 Klonenden konnten bestimmt werden, für die ein Treffer zu allen drei Genomen vorliegt. Für die Homologiesuche gegen Gerste wurden andere Sequenzparameter gewählt (`www.vmatch.de`, l=100, e=1) als für *Brachypodium* und Reis (l=100, seedlength=12, exdrop=3, identity=75). Die Anteile der Klonenden, die eine Zuordnung zu nur einem Genom aufweisen, sind gering: Für *Brachypodium* sind 26.0% der Treffer zu Klonenden spezifisch, während die Werte für Reis 15.4% sowie Gerste mit 15.7% geringer ausfallen. Es liegt vor allem daran, dass nur genetisch verankerte Gene, also nur 15,719 genutzt wurden, während in *Brachypodium* und für Reis der Großteil der Gene eine Position auf einem der fünf bzw. zehn Chromosome aufweist und nur eine sehr geringe Anzahl auf nicht verankerten Scaffolds annotiert wurde. In Gerste wurde nur eine Spleißvariante bestimmt und erklärt die geringe Anzahl von über Sequenzhomologie zugeordneten Genen im Vergleich zu Reis und *Brachypodium*. Betrachtet man die Klonenden mit einer Verankerung in allen drei Genomen, werden 7,159 Klone erreicht, die auf 1,892 FP contig verankert sind. Diese FP contig repräsentieren 1.98 Gb relativ zu den 3.9 Gb verankerten FP contigs. Abbildung 3.6 zeigt die gute Übereinstimmung zwischen den über Syntenie verankerten Klonenden im Vergleich zur IBSC [2012] Verankerung.

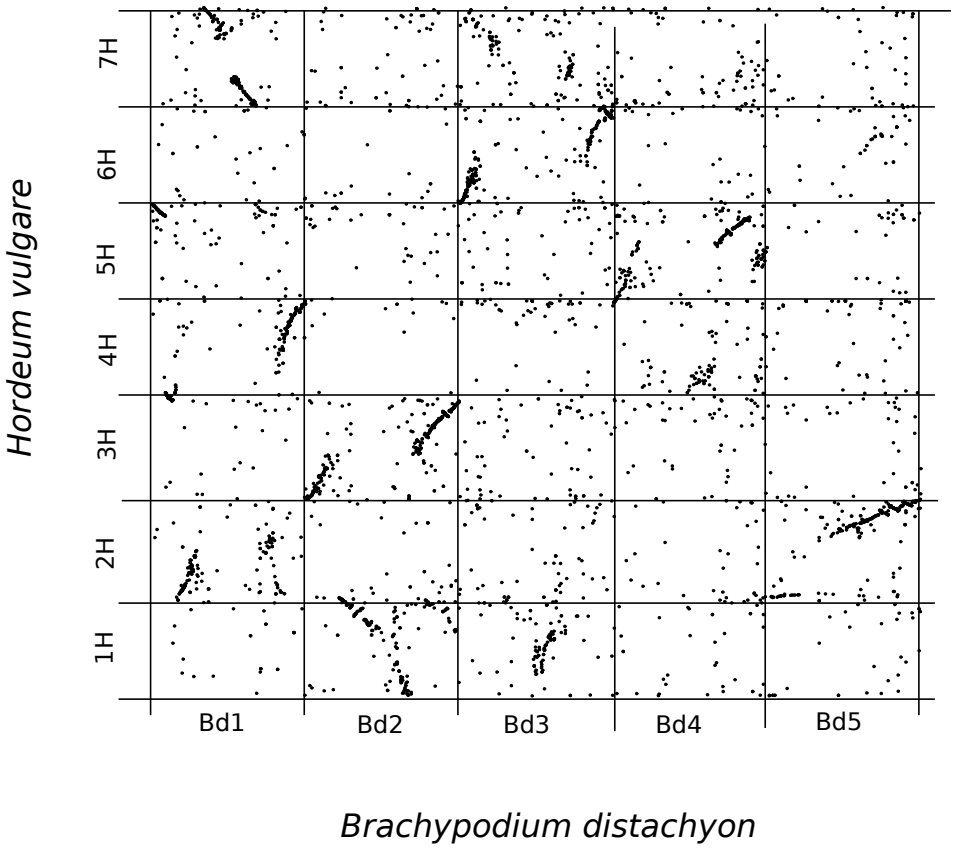

Abbildung 3.6: **Vergleich der syntenischen Verankerung über *BES* und Syntenie gegenüber den 3.9 Gb verankerten FP contigs aus IBSC [2012].** Die genomischen Positionen des *IBSC* sowie aus *Brachypodium* wurden normiert bezüglich ihrer Chromosomenlänge. 1H-7H ist Chromosom 1-7 in Gerste; Bd1-Bd5 entsprechen Chromosom 1-5 in *Brachypodium*.

3.1.9 Genverankerung in der Gerste

In der Arbeitsgruppe (Arbeit von Dr. Matthias Pfeifer und Manuel Spannagl) wurden Gene unter Einsatz von RNA-seq Sequenzen auf NGS contigs annotiert. Die daraus resultierenden Genmodelle wurden mit Gersten-Volllängen-cDNA zusammengefasst, um redundante Transkripte zu entfernen. 26,159 Gersten-Gene wurden bestimmt, 24,242 Gene mit Entsprechung zu einem NGS contig des Kultivars "Morex". Unterschiedliche Verankerungsstrategien wurden entwickelt, um Gene genetisch zu verankern (Abbildung 3.7): Direkt einem FP contig zugeordnete Sequenzen stellen die erste

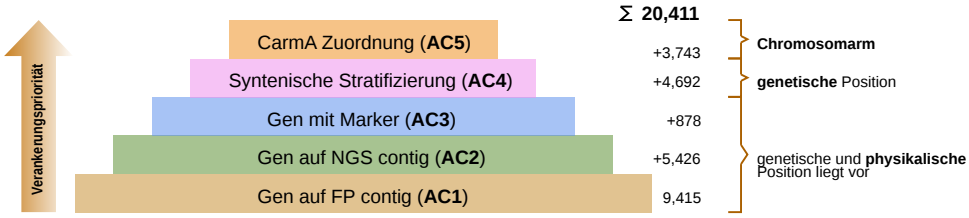

Abbildung 3.7: **Prioritäten bei der Verankerung von Genen in Gerste**. Die Richtung des Pfeils gibt die Priorität der Verankerung an beginnend mit *AC1* bis *AC5, anchoring class 5.*

Prioritätsstufe der Verankerung (AC1, *anchoring class 1*) dar. 9,415 Gerstengene liegen bereits auf verankerten FP contigs. Durch verankerte NGS contigs des Kultivars "Morex" kommen weitere 5,426 hinzu (AC2, *anchoring class 2*). Für diese NGS contigs liegt kein FP contig vor, allerdings eine eindeutige Markerverankerung. Zusätzlich steht die Markerzuordnung nicht im Widerspruch zur *CarmA*-Zuordnung. Kodierende Gene werden über Marker einer genetischen Position zugeordnet (AC3, *anchoring class 3*). Durch diesen Vorgang erfuhren weitere 878 Gene eine genetische Position. Für 15,719 Gene liegt somit eine genetische Position auf dem Genom vor (AC1-AC3). Über syntenische Stratifizierung (Kapitel 3.1.7) werden weitere 4,692 (AC4, *anchoring class 4*) Gene verankert. Insgesamt weisen 20,411 der 26,159 Gene eine genetische Position oder physikalische Position auf. Über *CarmA* und damit einem Chromosomenarm zugewiesene Gene bilden Prioritätsstufe 5 (AC5, *anchoring class 5*). In Summe können 24,154 oder 92.3% aller HC Gene in Chromosomenarme sortiert oder genetisch positioniert werden. Für positionierte Gene liegt neben der genetischen Position auch die ungefähre genomische Position, resultierend aus den kumulativen FP contig Längen, vor.

3.1.10 Validierung der Verankerung

Mit Hilfe von FPC (Soderlund et al. [1997]) wurden bei einem Sulston Score von e^{-45} Klone zu FP contigs zusammengefasst. Der Sulston Score ist ein Maß für die Wahrscheinlichkeit, dass Banden zwischen zwei Klonen übereinstimmen und damit Klone überlappen (Meyers et al. [2004]). Eine weitere Assemblierung mit e^{-25} erfolgte, um 2,000 potentiell überlappende FP contig-Paare zu bestimmen (Liste bereitgestellt und gefiltert von Dr. Ruvini Ariyadasa und Dr. Burkhard Steuernagel). Durch die 4,556 verankerten FP contigs wird überprüft, ob FP contigs, die in der weniger stringenten Assemblierung zusammengefasst wurden, auch in der stringenteren und für das *gene-ome* genutzten Version ähnliche genetische Positionen aufweisen. Eine Abweichung von 5 cM ist noch erklärbar über die Assoziation zu physikalischen Contigs in rekombinationsreichen Regionen. Die geforderten Unterschiede in der genetischen Position sollten demnach wenig restriktiv sein, weil durch die Integrationen von unterschiedlichen Markerkarten auch Abweichungen von einigen centiMorgan möglich sind. 80% der potentiell überlappenden Paare, für die beide FP contig eine genetische Position aufweisen, liegen innerhalb von $<$ 5 cM, rund 85% weisen eine Distanz von $<$10 cM auf, 90% liegen in einem 20 cM Bereich. Die Ergebnisse lassen vermuten, dass der maximale Fehler nicht nur auf einem Chromosomenarm, sondern einen deutlich kleineren Bereich beschränkt werden kann. Es liegt nahe, dass der Großteil der FP contig-Paare sehr wahrscheinlich zusammengefasst werden könnte und die genetische Verankerung der physikalischen Karte im hohen Maße korrekte Ergebnisse liefern sollte. Allerdings überlappen nicht alle dieser FP contigs, denn bei niedriger Stringenz führt FPC (Soderlund et al. [1997]) zu einer höheren Anzahl an schimärischen FP contigs (Philippe et al. [2012]). Von 2,000 Kandidatenpaaren, die potentiell überlappen, liegt eine *CarmA*-Zuordnung für 1,800 FP contig-Paare vor. 1,500 (83%) dieser Paare weisen auf denselben Chromosomenarm, 300 (17%) auf unterschiedliche, 200 (11%) FP contig Paaren fehlt zumindest eine Chromosomenarmzuordnungen, wodurch ein FP contig nicht genetisch positioniert wurde.

3.1.11 Das Gerste *gene-ome*

Das Gerste *gene-ome* repräsentiert die verankerten und mit Sequenzen ausgestatteten FP contigs, dargestellt in Abbildung 3.8. Insgesamt wurden 3.9 Gb der 4.99 Gb kumulativen FP contig-Länge verankert, rund 4.5 Gb konnten einem Chromosomenarm zugewiesen werden. Der Großteil der 26,159

HC Gene konnte genetisch oder über Syntenie verankert werden. Abbildung 3.8 stellt die verankerten FP contigs dar mit darauf zugewiesenen Sequenzen (sequenzierte Klonenden, sequenzierte Klone, NGS contigs). Weil NGS contigs und assemblierte Klonsequenzen den hoch-repetitiven Bereich nur unzureichend erklären, wurde die Verteilung der repetitiven Elemente auf Basis von Klonenden bestimmt. Die Annotation der repetitiven Elemente wurde von Dr. Heidrun Gundlach durchgeführt und Klonenden schließlich durch annotierte und manuell kuratierte repetitive Elemente maskiert. Anschließend wurde der durchschnittliche Grad an repetitiven Elementen berechnet und als *heat map* dargestellt. Die Chromosomenlänge wird durch Aneinanderreihung der verankerten FP contigs erreicht. Sequenzierte Klone werden pro 10 Mb-Bereich dargestellt und der Vergleich der genetischen mit physikalischer Distanz pro FP contig dargestellt.

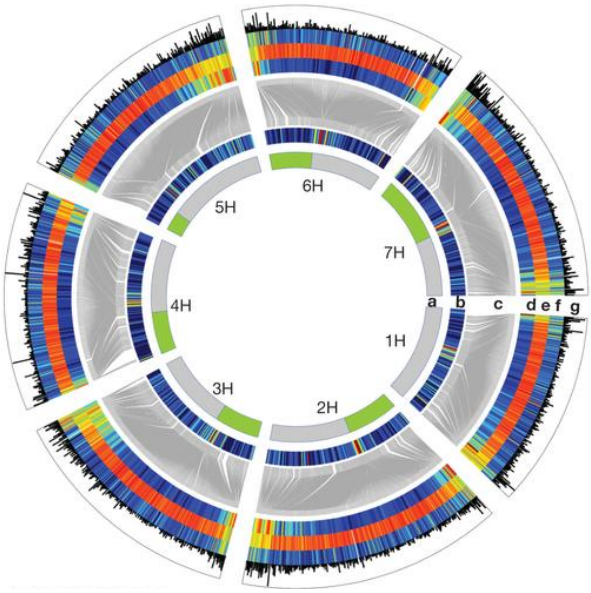

Abbildung 3.8: **Darstellung der genetisch verankerten FP contigs, das Gerste *gene-ome***. Heterogene Datenressourcen entlang der verankerten FP contigs. Entnommen aus IBSC [2012]. Spur a) zeigt den Vergleich gegen die in Chromosomenarme sortierten Sequenzen. Es zeigt den kurzen Arm, dargestellt in Grün im Vergleich zum langen Arm, dargestellt in Grau. Spur b)-d) zeigt die genetische und physikalische Position verankerter Gersten HC Gene. Das genetische Zentromer erstreckt sich über den Großteil der verankerten physikalischen FP contigs. Spur d) zeigt die Dichte an verankerten Genen, während Spuren e-f) die Verteilung der repetitiven Elemente zeigt. *LTR*-Retroelemente weisen eine Antikorrelation zu Genen auf. Die höchsten Werte liegen im Peri-zentromer und Zentromer. Spur g) zeigt die Verteilung der sequenzierten Klone. Der Großteil der Klone liegt in den distalen und proximalen Enden der Chromosomen.

3.1.12 Vergleich des Gerste *gene-ome* mit *Aegilops tauschii*

Im Jahr 2013 wurden die nächsten Verwandten der Weizenvorläufergenome des A- und D-Subgenoms von hexaploidem Weizen veröffentlicht (Jia et al. [2013], Ling et al. [2013]). Der nächste Verwandte des Vorläufergenoms des Weizen A-Genoms ist *Triticum urartu*, und *Aegilops tauschii* stellt das Vorläufergenom des Weizen D-Genoms dar. Für *Aegilops tauschii* wurde eine große Anzahl an Genen einer genetische Position zugeordnet und ermöglichte die Ergebnisse des Gerste *gene-ome* gegen *Aegilops tauschii* zu vergleichen. Tabelle 3.12 gibt einen Überblick über die Anzahl an verankerten Genen.

Gerste	*Ae. tauschii*	Gerstengene	*Ae. tauschii* Gene	Genpaare
1H	1D	1,532	2,082	992
2H	2D	1,845	2,664	1137
3H	3D	1,649	2,478	1027
4H	4D	972	1,749	666
5H	5D	1,730	2,553	1118
6H	6D	1,130	1,868	669
7H	7D	1,599	2,325	919
1H-7H	1D-7D	10,457	15,719	5,528

Tabelle 3.12: **Vergleich orthologer Gene zwischen Gerste und *Aegilops tauschii***. Die Zahlen stellen die Genpaare dar, die miteinander über den besten bidirektionalen Treffer zugeordnet sind und gleichzeitig eine genetische Position aufwiesen.

5,528 Genpaare aus Gerste und *Aegilops tauschii* wurden über den besten bidirektionalen Treffer zugeordnet (Mindesttrefferlänge von 75%, >30 Aminosäuren). Die Anzahl stellt 52.8% aller verankerten Gene des D-Vorläufergenoms dar. Der Vergleich der Syntenie über alle Gerstenchromosomen ist in Abbildung 3.10 angeführt sowie die Verteilung der Trefferlängen (Abbildung 3.9) und Sequenzidentität der Genpaare mit durchschnittlichen Trefferlängen von 300 Aminosäuren und einer durchschnittlichen Sequenzidentität von 90%. Tabelle 3.13 führt die durchschnittlichen Trefferlängen der orthologen Gene zwischen *Aegilops tauschii* und Gerste und die durchschnittlichen Sequenzidentität an. Die Sequenzidentitäten der Chromosomenvergleiche sind annähernd uniform mit Werten zwischen 89.9 (6H) und 91.5 (5H).

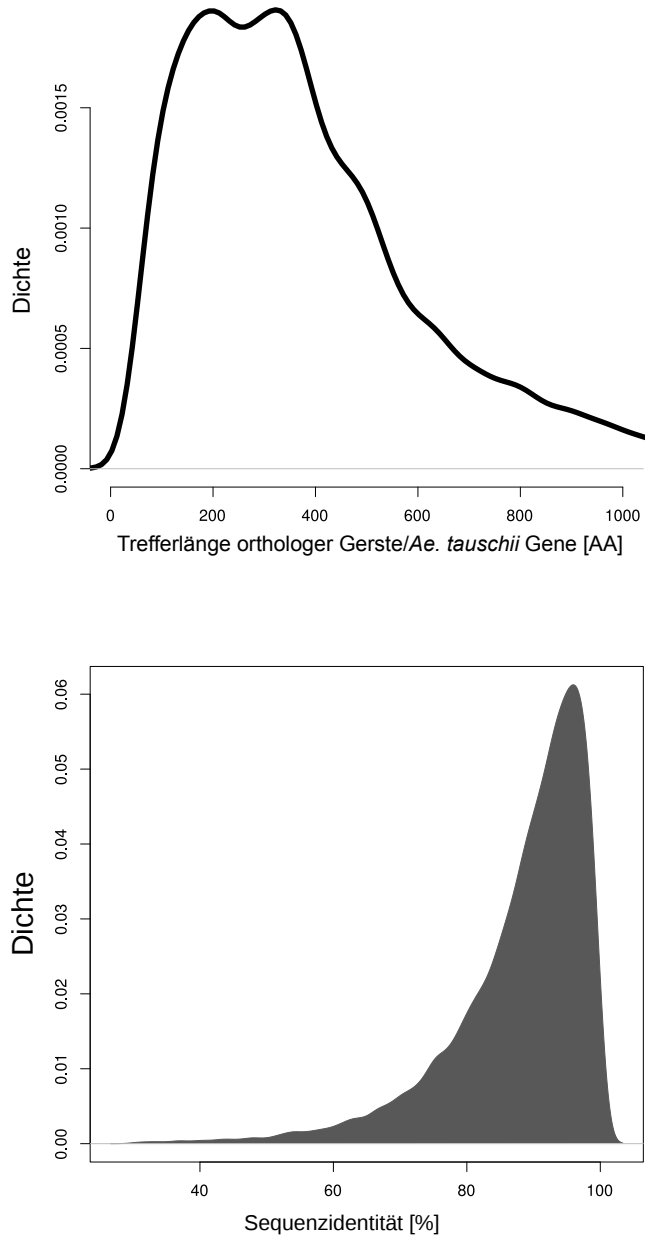

Abbildung 3.9: **Trefferlänge und Sequenzidentität in orthologen Genen zwischen *Aegilops tauschii* und Gerste.**

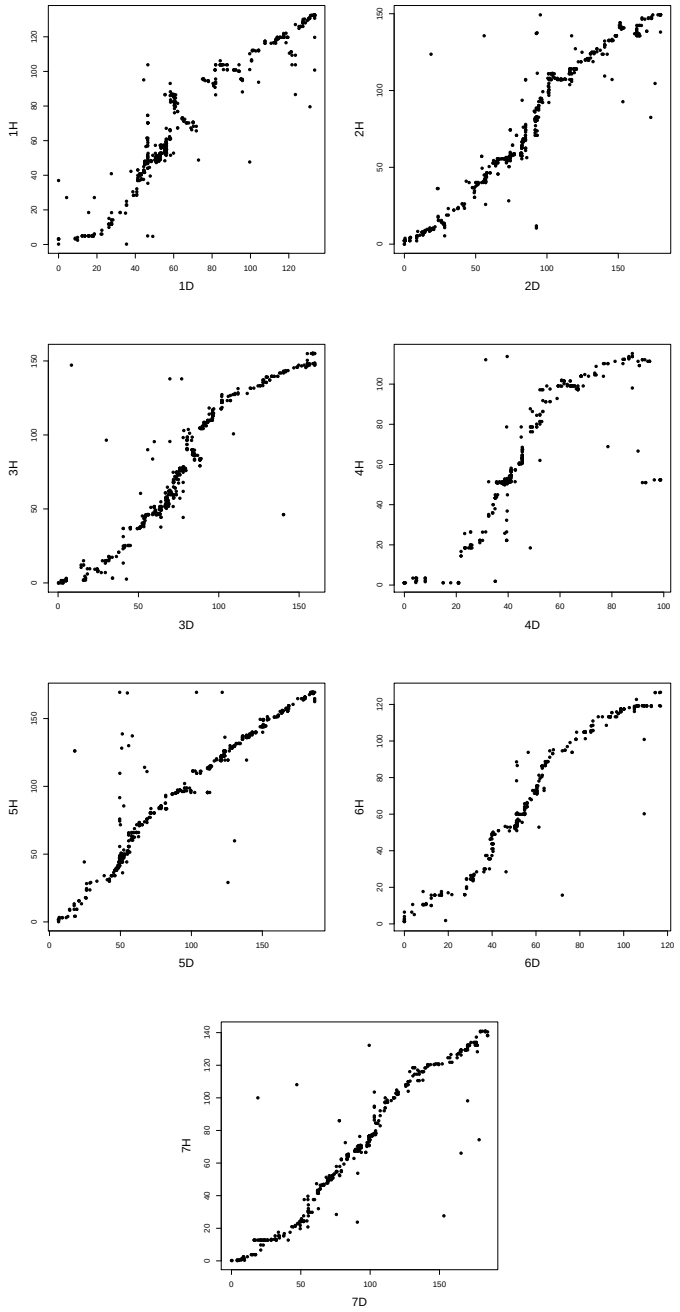

Abbildung 3.10: **Chromosomen-Vergleiche zwischen *Aegilops tauschii* und Gerste**. Vergleiche zeigen eine starke ausgeprägte Kollinearität mit Ausnahme einer größeren Inversion auf 1H/1D.

Chromosom	Genpaare	Länge	Sequenzidentität
1H	992	345.0	90.5
2H	1,137	357.0	90.5
3H	1,027	347.0	90.4
4H	666	335.0	91.3
5H	1,118	343.0	91.5
6H	669	361.0	89.9
7H	919	340.0	90.2

Tabelle 3.13: **Sequenzidentität zwischen Gerste und *Aegilops tauschii*.** Angabe der Länge in Anzahl an Aminosäuren (AA).

3.2 Verankerung von Weizenchromosom 6A

Für die Gerste wurden Klone mit Hilfe der *High-Information Content Fingerprinting*-Technologie (*HICF* Technologie; (Ding et al. [2001])) zu FP contigs assembliert. Für das Weizenchromosom 6A wurde die neu entwickelte "Whole Genome Profiling" (WGP^{TM})-Technologie (van Oeveren et al. [2011])) verwendet und eine Klonbibliothek für jeden Chromosomenarm einzeln angelegt. Der Restriktionsverdau erfolgte durch *Hin*dIII und generierte 49,152 Klone für den kurzen Arm und 55,296 Klone für den langen Arm. 22,656 Klone für den kurzen und 24,575 Klone für den langen Arm wurden zu FP contigs assembliert. Die physikalische Karte wies eine Genomabdeckung von $8x$ auf, im Vergleich zu einer $16x$ Genomabdeckung aller Klone der beiden Klonbibliotheken. Für Klone wurden informative WGP^{TM}-*Tags* nach Verdau durch zwei Restriktionsenzyme (*Hin*dIII, *Mbo*1) bestimmt und die Restriktionsschnittstellen umgrenzenden genomischen Bereiche mit einer Länge von 100 bp sequenziert. Die *Tags* dienten dazu, um Klone in FP contigs zu assemblieren und wurden anschließend als Verbindungsstück der physikalischen Karte zu den deutlich längeren NGS contigs des *IWGSC* genutzt. Die Assemblierung der Klone zu FP contigs wurde mit FPC (Soderlund et al. [1997]) und LTC (Frenkel et al. [2010]) durchgeführt. Die Integration der unterschiedlichen Sequenzressourcen ist in Abbildung 3.11 dargestellt.
Im Rahmen dieser Arbeit sollten die Performanz beider Programme unter unterschiedlichen Stringenzkriterien verglichen werden. Anschließend sollte die beste Assemblierung (LTC oder FPC) für die Verankerung und Aneinanderreihung der FP contigs aus Chromosom 6A ausgewählt werden und um

diesen weiteren Sequenzen zuzuweisen. Sequenzierte *Tag*-Sequenzen wurde durch NGS contigs (IWGSC [2014]) erweitert und NGS contigs gegen die deutlich längeren Sequenzen aus den Vorläufergenomen von *Triticum urartu* (Ling et al. [2013]) und *Aegilops tauschii* (Jia et al. [2013]) verglichen und bei starker Sequenzhomologie dem FP contig zugewiesen (siehe Kapitel 2.4). Es wurde außerdem untersucht, ob *Triticum urartu* contigs eine Sequenzhomologie zu weiteren, bisher nicht berücksichtigten NGS contigs in Weizen aufwiesen und diese NGS contigs ebenfalls dem FP contig zugeordnet. Die längeren Scaffolds in *Triticum urartu* wurden genutzt, um potentielle überlappende FP contigs zusammenzufassen, wenn ein Ende eines *Triticum urartu* Scaffolds durch den einen FP contig, das andere Ende des Scaffolds einen anderen FP contig abgedeckt wurde. *Triticum urartu* wurde für diese Analyse genutzt, weil es der Vorfahr des A-Subgenoms und damit von Weizenchromosom 6A ist. Das D-Genom von Weizen leitet sich von *Aegilops tauschii* ab und enthält sehr wahrscheinlich eine deutlich größere Anzahl an strukturellen Unterschieden. Die FP contigs mit allen darauf verankerten Sequenzen wurden anschließend durch publizierte Markerkarten (Saintenac et al. [2013], Poland et al. [2012]) genetisch verankert und FP contigs entlang der integrierten genetischen Karte angeordnet. Die Integration der genetischen Karten und die Erstellung der verankerten physikalischen Karte in Weizenchromosom 6A wird in Kapitel 2.4 beschrieben.

3.2.1 Vergleich von LTC und FPC im Weizenchromosom 6A

Einer der Schwerpunkte der 6A-Studie lag im Vergleich der beiden Klonassemblierungsprogramme LTC (Frenkel et al. [2010]) und FPC (Soderlund et al. [1997]). Mit FPC wurden dreizehn Assemblierungen von Klonen zu FP contigs (Contigs) durchgeführt, bei Signifikanzen zwischen e^{-75} und e^{-11}. Diese Assemblierungen wurden in LTC bei einem e-value von $1e^{-10}$ gegenübergestellt. Im Weizenchromosomenarm 6AS wurden mit LTC 1,214 FP contigs assembliert, im Vergleich zu 539 ($1e^{-11}$) bis 559 Contigs ($1e^{-75}$), die mit FPC assembliert wurden. Zwischen 44-46% aller Contigs waren exklusiv für LTC. Kein Klon dieser Contigs wurde in FPC berücksichtigt. Diese exklusiven Contigs wiesen im Durchschnitt 2.6 Klone auf, im Gegensatz zu den 3.5 Klonen in Contigs mit gleicher Klonkomposition in beiden Programmen. 1-15% ($1e^{-75}$ bzw. $1e^{-11}$) der Contigs in LTC wiesen eine geringere Klonanzahl als FPC auf, während LTC-contigs mit einer größeren Klonanzahl in LTC 9-13% aller Contigs stellten. Die besten Übereinstimmung von LTC

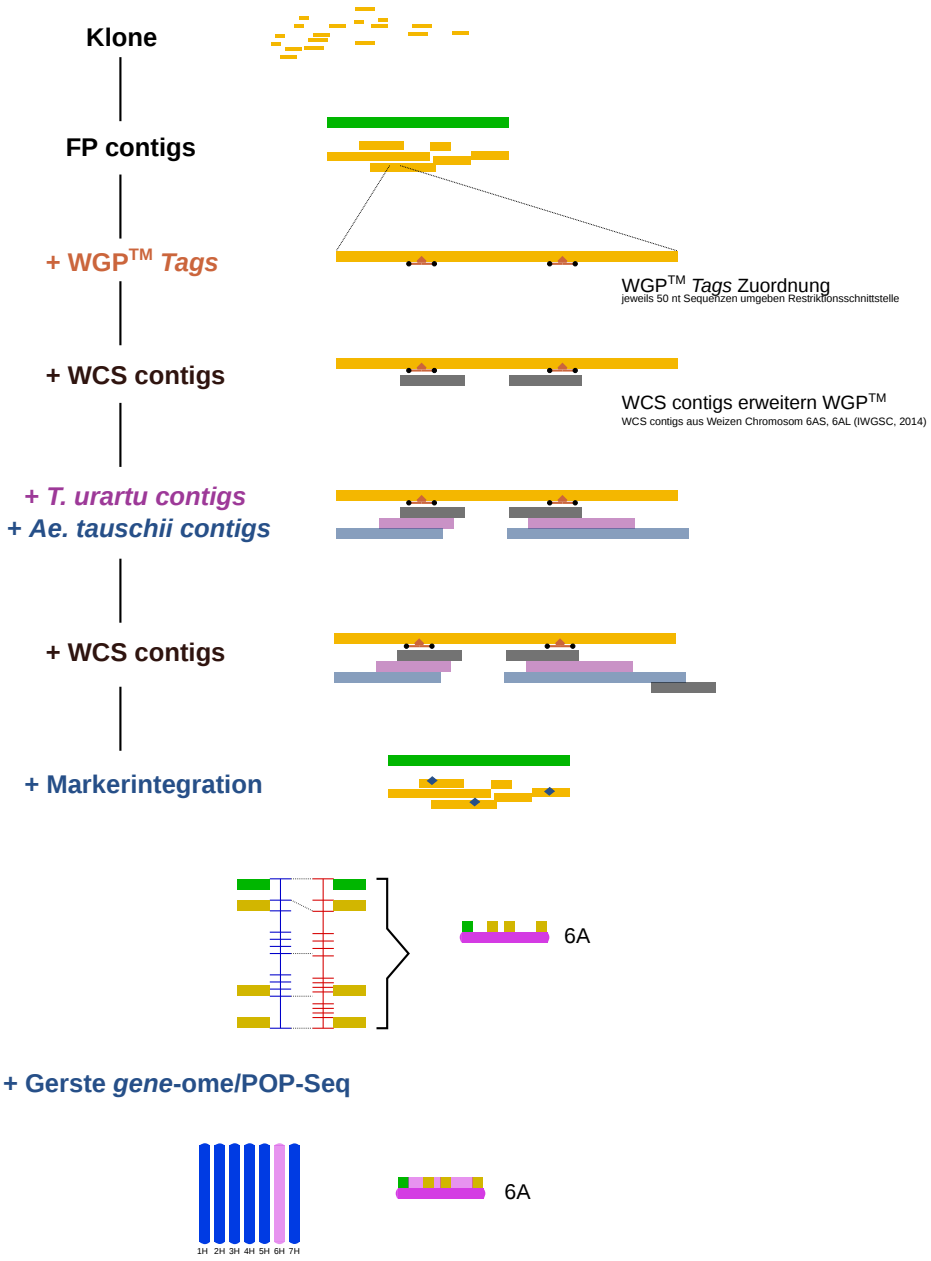

Abbildung 3.11: **Integration einzelner Sequenzressourcen in Weizenchromosom 6A**.
WGP^{TM} *Tags* werden über Sequenzhomologie mit NGS contigs verbunden. Diese werden durch *Contigs* aus *Aegilops tauschii* und *Triticum urartu* erweitert und genutzt, um FP contigs mit Hilfe von *in silico* Markern einer genetischen Position zuzuweisen. Anschließend werden FP contigs ohne genetische Position mit verankerten Ressourcen aus Gerste positionell verankert.

und FPC erfolgte mit den Stringenzkriterien $1e^{-11}$ bis $1e^{-20}$, während mit $1e^{-75}$ zwar eine etwas größere Anzahl an Contigs (559) assembliert wurde, aber nur 20% der Contigs in LTC eine gleiche Klonkomposition wie FPC aufwiesen. Im Weizenchromosomenarm 6AL wurden mit LTC 1,108 FP contigs bestimmt, im Vergleich zu 368 ($1e^{-11}$) bis 389 ($1e^{-75}$) Contigs, die mit FPC assembliert wurden. Rund ein Drittel der LTC-contigs waren spezifisch für diese Assemblierung, während 12.6% ($1e^{-11}$) der LTC-contigs eine gleiche Klonkomposition wie FPC aufwiesen, bei $1e^{-75}$, hingegen nur mehr 10.2%. Bereits in Frenkel et al. [2010] wurde LTC als stabileres Programm für Weizen beschrieben. Aus diesem Grund und weil LTC eine höhere Klonanzahl zu FP contig assemblieren konnte, wurde schließlich LTC ausgewählt, um das Weizenchromosom 6A funktionell und strukturell zu beschreiben.

3.2.2 Genetische Verankerung der FP contigs

Für die Verankerung von Weizenchromosom 6A wurden zwei *in silico* Markerkarten genutzt (Saintenac et al. [2013], Poland et al. [2012]) und auf die mit Sequenzen angereicherte physikalische Karte übertragen (siehe Kapitel 2.4). In Saintenac et al. [2013] wurde eine *GBS*-basierte Karte erstellt und ein Array auf Basis von 9,000 SNPs als Referenzkarte genutzt und 421,065 Marker bereitgestellt. In Poland et al. [2012] wurden 240,000 Sequenzen bestimmt und ebenfalls die *GBS*-Technologie eingesetzt. Mit Markern der Saintenac et al. [2013] Karte wurden ≈100 Mb der physikalischen Karte pro Chromosomenarm einer genetischen Position zugewiesen, während durch Marker der Poland et al. [2012] Karte, 363 Mb auf dem kurzen Chromosomenarm (6AS) sowie 383 Mb auf dem langen Chromosomenarm (6AL) verankert wurden (Tabelle 3.14). Von insgesamt 13,683 *GBS* Markern der (Saintenac et al. [2013] Karte wurden 3,025 Marker auf 6AS und 2,393 Marker auf 6AL verankert, zusammen 39.6% aller Markersequenzen dieser Karte (Abbildung 3.12). Die kurzen Längen der Marker und die erforderliche Stringenz bei der Zuordnung auf die genomische Sequenz, sowie die Assemblierung vorwiegend genreicher genomischer Bereiche sind Gründe für den hohen Anteil an Markern ohne Zuordnung. Es konnten mit einer leicht adaptierten Verankerungsstrategie wie in Gerste (siehe Kapitel 3.1.4) 662 Mb FP contigs verankert werden. Bei einer kumulativen Länge von 1,048 Mb für alle FP contigs konnte ein Verankerungsgrad von 65% erreicht werden.

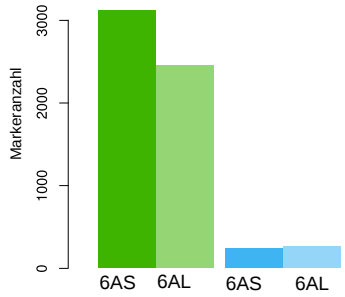

Abbildung 3.12: **Anteil an Markern an der genetischen Positionierung von FP contigs in Weizenchromosom 6A**. Die Marker entstammen aus Poland et al. [2012] (dargestellt in Grün) und aus Saintenac et al. [2013] (dargestellt in Blau).

Arm	Markerressource	verankert	insgesamt
6AS	Saintenac et al. [2013]	92	522
6AL	Saintenac et al. [2013]	91	543
6AS	Poland et al. [2012]	363	522
6AL	Poland et al. [2012]	383	543

Tabelle 3.14: **Anteil an Markern an der genetischen Positionierung von FP contigs in Weizenchromosom 6A**.
Die kumulativen Längen der physikalischen Karten beider Chromosomenarme sowie deren verankerter Anteil. Alle Angaben in Mb.

3.2.3 Gene auf Weizenchromosom 6A

Durch das Internationale Weizengenomkonsortium (IWGSC [2014]) wurden 124,201 Gene bestimmt, indem RNA-seq Datensätze aus fünf unterschiedlichen Gewebetypen und unter Verwendung von Volllängen-cDNAs auf die in Chromosomenarme sortierten NGS contigs übertragen wurden. Die potentiellen Genkandidaten wurden anschließend gegen Referenzgenome wie Reis, *Sorghum*, *Brachypodium* und Gerste verglichen. Gene wurden in mehrere Konfidenzklassen unterteilt: Gene der 1. Klasse wiesen eine Abdeckung von mehr als 70% zu einem Referenzgen auf. Gene der 2. Klasse wiesen eine Abdeckung zwischen 50% und 70% auf, Gene der 3. Klasse eine Abdeckung von 30% und 50% und Gene der 4. Klasse eine Abdeckung von >0% bis 30% auf. Durch die Integration der einzelnen Datenressourcen wurden neben den verankerten WCS contigs auch den darauf annotierten HC (*high-confidence*) Genen eine genetischen Position zugewiesen. Abbildung 3.13 zeigt den Anteil der Gene für Chromosom 6A für die einzelnen Konfidenzklassen: 69% der HC1 Gene sind der physikalischen Karte zugeordnet, während 57-61% der

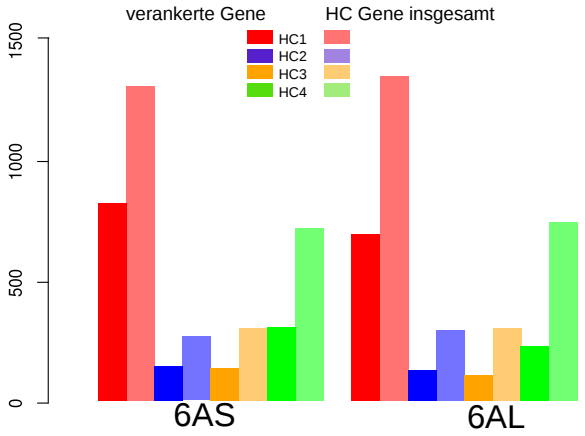

Abbildung 3.13: **Anteil an verankerten Genen in Weizenchromosom 6A**. Verankerte und gesamte Genanzahl in Weizenchromosom 6A aufgetrennt in die vier Genkonfidenzklassen 1-4 (HC1-HC4).

HC2-HC4 auf einem genetisch verankerten FP contig liegen (Poursarebani et al. [2014]). Der relative Anteil nimmt von der höchsten zur niedrigsten Konfidenzklasse ab, weil Gene in niedrigeren Konfidenzklassen kürzerere Sequenzen haben und entsprechend seltener auf einen genetischen Marker oder WGP^{TM}-*Tag* treffen. In Summe wurden für Weizenchromosom 6A 3,843 Genmodelle verankert.

3.2.4 Gerste zur Verankerung weiterer FP contigs

Genetisch verankerte Ressourcen aus der Gerste wurden genutzt, um in Weizenchromosom 6A zusätzliche und bisher nicht verankerte FP contigs einer genetischen Position zuzuweisen. Nach der Publikation des Gerste *gene-omes* konnten in einem auf Populationsgenetik basierenden Ansatz (einem Vorgehen, das als POPSEQ bezeichnet wurde) rund 1.22 Gb des 1.8 Gb umfassenden Gersten Assemblierungsdatensatzes genetisch positioniert werden (Mascher et al. [2013]). Aufgrund der syntenischen Konservierung von Gerste und Weizen und vor allem aufgrund der vielen genetisch verankerten Sequenzen in Gerste wurden POPSEQ verankerte NGS contigs ebenfalls genutzt, um verbliebene 6A FP contigs genetisch zu positionieren. Zunächst wurde überprüft, ob die Gersten- und Weizenverankerungen tatsächlich eine ähnliche genetische Verankerung für einen Weizen FP contig bereitstellen. Dazu wurden die FP contig-Verankerungen der Gersten-Verankerung des *IBSC*

und von POPSEQ gegen die Weizen Verankerung verglichen (siehe Kapitel 2.4). 422 FP contigs aus Chromosom 6A wiesen eine Verankerungsposition in allen drei Verankerungsstrategien auf (POPSEQ, *IBSC*, Chromosom 6A) mit paarweisen Korrelationskoeffizienten von >0.9. Die Gerste ist somit ein gutes strukturelles Modell für Weizen und legitimierte die Verankerung weiterer 101 FP contigs (53 Mb) durch POPSEQ und damit weiterer vier FP contigs (3 Mb) mittels Gerstengenen (N=15,719).

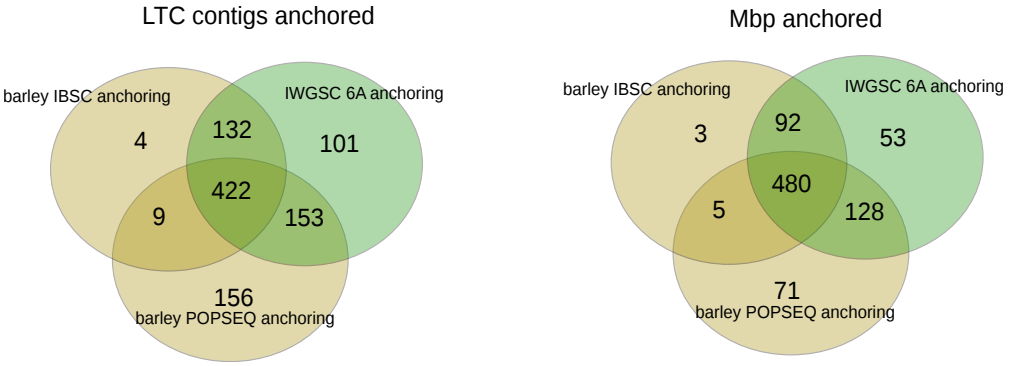

Abbildung 3.14: **Verankerungsstrategie der *in silico* Verankerung von Weizenchromosom 6A**. Einfluss der verschiedenen Gersten- und Weizenressourcen auf die Verankerung von FP contigs. Entnommen aus Poursarebani et al. [2014].

Abbildung 3.14 zeigt den Anteil an FP contigs sowie Anzahl an Mb der mit LTC-erstellten FP contigs mit Hilfe der Marker-basierten Verankerung und im Vergleich zu den über Gerste ermöglichten Verankerungen durch *IBSC* und POPSEQ. In Summe konnten unter Zuhilfenahme von Gerste, 831 Mb der 1,065 Mb (kumulative Gesamtlänge der FP contigs) des Weizenchromosoms 6A und 76.4% der Gene, die im IWGSC [2014] diesem Chromosom zugewiesen wurden (N=5,024), genetisch verankert werden.
Das verankerte Weizenchromosom 6A zeigte, wie gut nahe verwandte Genome dazu geeignet sind, um Sequenzinformationen zu übertragen. Das verankerte Weizenchromosom wurde dazu in 20 Mb große und nicht-überlappende Bereiche unterteilt und der verankerte genomische Anteil (in bp) für *Triticum urartu*, *Ae. tauschii* und *Triticum aestivum* bestimmt (Abbildung 3.15). An den distalen und proximalen Enden der Chromosomen, wie durch den vergleichsweise geringen Anteil an repetitiven Elementen zu erwarten, liegt die größte Sequenzmenge vor. Die *Tag*-Sequenzen sind annähernd gleichverteilt über das gesamte Chromosom, eine Notwendigkeit, damit die physikalische Karte das Genom abdeckt. Die geringere Anzahl von repetitiven Elementen an den Telomer-nahen Bereiche erlaubte außerdem eine deutlich

größere Datenmenge aus den Vorläufergenomen zu verankern. Aufgrund der kürzeren evolutionären Distanz von Weizenchromosom 6A zu *Triticum urartu* wurden außerdem im Vergleich zu *Aegilops tauschii* annähernd doppelt so viele NGS contigs der physikalischen Karte auf.

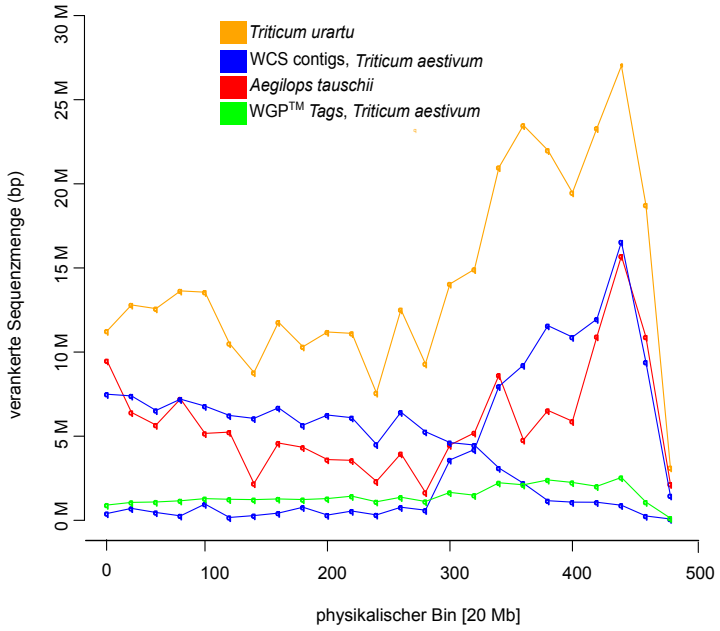

Abbildung 3.15: **Zusätzliche Verankerung von Weizen FP contigs durch nah-verwandte Weizengenome**. Anteil an Sequenzen der Vorläufergenome und Genom eigener Sequenz auf der verankerten physikalischen Karte. Adaptiert nach Poursarebani et al. [2014].

3.3 Verankerung von Weizenchromosomen 1D, 4D und 6D

Mit Hilfe der Durchflusszytometrie wurden 21 Weichweizenchromosomen aufgrund eines ähnlichen DNA-Gehalts in vier Gruppen unterteilt (Vrana et al. [2000]). Eine dieser Gruppen (Gruppe I) wurde durch die Chromosomen 1D, 4D, 6D und 7D definiert und später konnten die drei Chromosomen 1D, 4D und 6D, ohne 7D für die Erstellung von BAC-Bibliotheken genutzt werden. Das genomische Material der drei Chromosomen 1D, 4D und 6D wurde genutzt, um Klonbibliotheken durch Verdau mit dem Restriktionsenzym *Hin*dIII anzulegen (Janda et al. [2004]). Die 280,000 Klone sollten zur Erstellung einer physikalischen Karte genutzt werden. Um die Integration der Daten zu erlauben, wurden für die Weizenchromosomen 1D, 4D und 6D $\approx\frac{1}{3}$ der Klone des minimalen überspannenden Pfads sequenziert und im Zuge dieser Arbeit assembliert, genetisch verankert und gegen die WCS contigs des *IWGSC* verglichen. Diese Sequenzen eines Klons sind einem Sequenzpool angeordnet, der jeweils 384 Klone umfasst. 59 Pools in Summe stellen die drei Chromosomen dar und beinhalten in Summe 7,064 Klone. Ein Klon weist jeweils eine x-, y- oder z-Koordinate auf, eine Achse wird jeweils von einem bestimmten Pool ausgefüllt (Abbildung 3.16). Die Schnittmenge der drei Achsen definiert einen Klon (Abbildung 3.16). Die einzelnen Pools wiesen Unterschiede in der Sequenzmenge aus, resultierend aus Kontaminationen mit *E.coli* und durch mitgeführte Vektorsequenzen. Die gereinigten Sequenzen wurden mit SOAPdenovo (Luo et al. [2012]) assembliert und Sequenzen aller 59 Pools berücksichtigt (siehe Kapitel 2.5). Im Rahmen dieser Arbeit sollten die Sequenzen den einzelnen Klonen zugeordnet werden und mit Hilfe der FP contig-Karte zur genetischen Verankerung der FP contigs genutzt werden. Die Assemblierung der Klone in FP contigs wurde analog zum bereits beschriebenen Ansatz in Gerste für Weizenchromosome 6A mit FPC (Soderlund et al. [1997]) durchgeführt. Die einzelnen Schritte der Integration der unterschiedlichen Daten für Weizenchromosomen 1D, 4D und 6D sind in Abbildung 3.16 schematisch dargestellt.

3.3.1 Dekonvolution und genetische Verankerung

Die Vorgehensweise, die eine Konsensussequenz einem bestimmten Klon zuordnet, wird als “Dekonvolution” bezeichnet. Sequenzen der 59 Pools wurden mit Hilfe von SOAPdenovo assembliert (Luo et al. [2012]). Anschließend

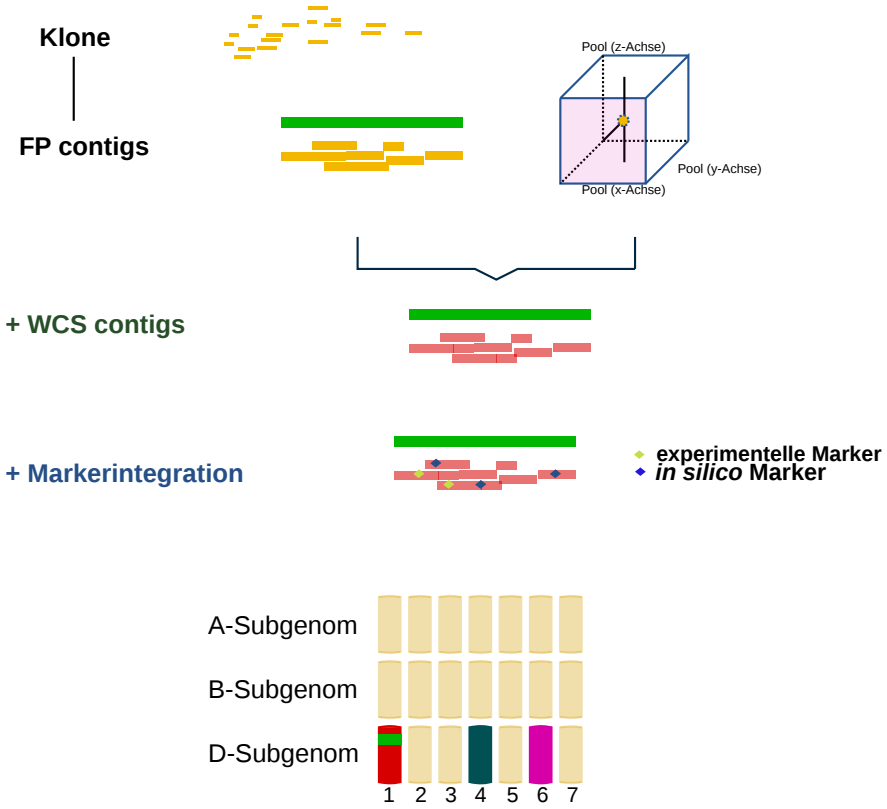

Abbildung 3.16: **Integration einzelner Sequenzressourcen in Weizenchromosomen 1D, 4D und 6D**. Klone werden durch FPC (Soderlund et al. [1997]) zu FP contigs assembliert. Anschließend werden die Sequenzen aus 3D-Klonpoolsequenzierung einzelnen Klonen zugeordnet und anschließend durch genetische Marker verankert.

Klonanzahl	Sequenzanzahl	Sequenzlänge	%
Kein Klon[$]	216,973	102,314,001	18.7%
Ein Klon	794,880	439,570,161	80.2%
Zwei Klone	14,480	4,636,538	0.8%
Mehrere Klone	4,266	1,698,164	0.3%
$\sum$	1,030,599	548,218,864	100.0%

Tabelle 3.15: **Ergebnisse der Datendekonvolution**. Die Assemblierung wurde mit Hilfe des Assemblierungsprogramms SOAPdenovo (Luo et al. [2012]) bestimmt, eine k-mer Größe von 85 wurde genutzt.
[$]Dekonvolution einer Sequenz zu einem bestimmten Klon war nicht möglich. Angabe der Länge in bp.

wurde jeder Contig gegen die Sequenzen aller Pools über Sequenzhomologie verglichen. Die drei Pools mit höchster Abdeckung auf diesen Sequenzen wurden vermerkt. Aufgrund der 3D-Struktur der Pools sollte ein Klon jeweils in einem Pool der x-, y- und z-Achse vorliegen (siehe Kapitel 2.5). Die Assemblierung wies eine Gesamtlänge von 548 Mb auf (Tabelle 3.15). Für 439 Mb an Sequenzinformation konnte einem bestimmten Klon zugewiesen werden, während für 102 Mb keine eindeutige Zuordnung zu einem bestimmten Klon möglich war. Die mit Sequenzen versehenen Klone wurden anschließend durch Dr. Sunish Sehgal bei einem e-value von e^{-60} zu FP contigs assembliert. Klone aus Weizen wurden zusammen mit Klonen aus *Aegilops tauschii* (Luo et al. [2013]) zu FP contigs assembliert und experimentelle Marker dieser Studie auch in Weizen genutzt. Dies wird dadurch ermöglicht, weil das D-Genom und *Aegilops tauschii* strukturell stark konserviert sind und eine evolutionäre Distanz von < 0.4 Millionen Jahre aufweisen (Marcussen et al. [2014]). Die Methodik *CarmA* wurde genutzt, um einen FP contig auf Basis aller Sequenzen aller Klone einem Chromosomenarm zuzuweisen. Schließlich wurden experimentelle Marker und zusätzlich in dieser Studie 2,450 SNPs und *Tags* der *GBS*-Technologie genutzt, um die physikalischen contigs genetisch zu verankern (siehe Kapitel 2.5). Für Weizenchromosom 1D wurden 160 Mb verankert, für Chromosom 4D 129 Mb und für Chromosom 6D 133 Mb. Für 1D weist die genetische Karte eine Länge von 163 cM auf, für 4D und 6D 171 und 221 cM gemessen.

3.4 Ährenfusariose in Weizen

3.4.1 Aufbau des Experiments

Bisher wurde dargestellt, wie die Verankerungen der physikalischen Karten in Gerste und in einzelnen Weizenchromosomen durchgeführt wurden. Diese mit Sequenzen angereicherten physikalischen Karten erlauben es, Kandidatengene positionell zu bestimmen, wenn größere genomische Fragmente von einigen hundert kb bis zu einigen Mb bereitstehen. Vorerst liegen diese Bereiche aber nur teilweise sequenziert vor.
Dieses Kapitel beschreibt, wie das Gerste *gene-ome* genutzt wurde, um die Resistenz von unterschiedlich resistenten Weizenlinien (Tabelle 3.16) gegenüber dem Pilz *F. graminearum* zu untersuchen. Der Umweg über Gerste war notwendig, weil zu diesem Zeitpunkt keine Genannotation für Weizen vorlag. Das Genom von Weizen wurde allerdings im Laufe der Studie basierend auf einem auf 454-Sequenzierung-basierten Assembly mit niedriger Genomabdeckung beschrieben (Brenchley et al. [2012]). Die Übertragung der Transkriptomsequenzen auf die Referenzsequenz, die Bestimmung von differentiell exprimierten Genen und die Analyse ausgewählter Genfamilien wurden im Rahmen dieser Arbeit durchgeführt und in der Arbeitsgruppe ein Ko-expressionsnetzwerk auf Basis der beschriebenen Linien (Tabelle 3.16) erstellt (Kugler et al. [2013]). Als Zeitpunkte wurden 30 und 50 Stunden

Weizenlinie	*Fhb1*	*Qfhs.ifa-5A*	Befallsgrad
CM-82036	+	+	sehr resistent
NIL1	+	+	stark resistent
NIL2	+	-	resistent
NIL3	-	+	resistent
NIL4	+	+	anfällig
Remus	-	-	sehr anfällig

Tabelle 3.16: **Elternpflanzen sowie nahezu isogene Linien und ihre Resistenzbereiche**. Weizenlinien und Auflistung welche QTLs (*Fhb1*, *Qfhs.ifa-5A*) diese besitzen. NIL = nahezu isogene Linien (NILs) auf Basis des FHB anfälligen rekurrenten Elter "Remus" unterscheiden sich in der Präsenz des jeweiligen QTLs.

nach Infektion gewählt. Beide Zeitpunkte stellen kritische Phasen in der Pflanzen-Pathogen Interaktion dar und umfassen den Wechsel vom biotrophen zum nekrotrophen Lebensstil des Pilzes sowie die damit einhergehende Bildung von Mykotoxinen. Neben der anfälligen Weizensorte "Remus"

und der resistenten Weizensorte “CM-82036” wurden außerdem vier nahezu isogene Linien untersucht, mit “Remus” als rekurrentem Elter. Genotypen wurden selektiert, die keinen, einen oder beide Resistenzbereiche aus der Weizensorte “CM-82036” aufwiesen, und mit NIL1-NIL4 bezeichnet (Tabelle 3.16). Sie stellen nahezu isogene Linien dar und besitzen keinen, einen oder beide Resistenzbereiche (*Fhb1*, *Qfhs.ifa-5A*), die einen unterschiedlichen Befallsgrad bewirken.

3.4.2 Übertragung der Expressionsdaten auf die genomische Weizenreferenz

In Brenchley et al. [2012] wurde eine genomische Assemblierung von Weizen erstellt, basierend auf einem Assembly aus Sequenzen der 454-Sequenzierung, bei 5-facher Genomabdeckung. Die Expressionsdaten aus den fünf unterschiedlich toleranten Weizenlinien wurden anschließend auf die genomische Referenzsequenz übertragen und Genmodelle und ihre Expression bestimmt (siehe Kapitel 2.6). Für jede Weizenlinie wurden Daten aus zwei Behandlungen (Negativkontrolle, Pilzbefall), aus zwei Zeitpunkten (30 und 50 Stunden nach Befall) sowie von drei biologischen Replikaten bereitgestellt, in Summe 1.8 Tb Expressionsdaten. 233,780 putative Genmodelle wurden mit Cufflinks (Trapnell et al. [2012]) bestimmt und die Sequenzhomologie zur Gerste genutzt, um aus dieser großen Anzahl an Gen-Kandidaten proteinkodierende Gene zu bestimmen. 15,360 (CM-82036) bis 15,797 (NIL1) Gerstengene aus den insgesamt 26,159 Gerstengenen wiesen Sequenzhomologie zu einem Weizen Genmodell auf (Tabelle 3.17). Im nächsten Schritt wurde auf Basis von Weizengenen mit besonders hoher Varianz in den untersuchten Bedingungen, ein Ko-expressionsnetzwerk erstellt (Kugler et al. [2013], Kapitel 2.6). Ein Ko-expressionsnetzwerk besteht aus Modulen, die Gruppen von Genen mit einem ähnlichen Expressionsprofil umfassen. Acht Module mit insgesamt 3,412 Genen wurden bestimmt (Abbildung 3.17). Zwei Module (Modul B, Modul G) zeigten eine unterschiedliche Expression zwischen Befall und Kontrolle. 1,148 Gene des Moduls B wiesen eine Überrepräsentierung von Begriffen der Gene-Ontology-Datenbank auf, wie *z. B.* Gluthatione S-transferasen, Glycosyltransferasen, Proteinkinasen und WRKY-Transkriptionsfaktoren. Die jeweiligen Genfamilien spielen eine Rolle in der Detoxifizierung des Mykotoxins Deoxynivalenol (DON) (Poppenberger et al. [2003]) oder sind an der Signaltransduktion bzw. der Expressionskontrolle (*z. B.* WRKY Transkriptionsfaktoren) von möglichen

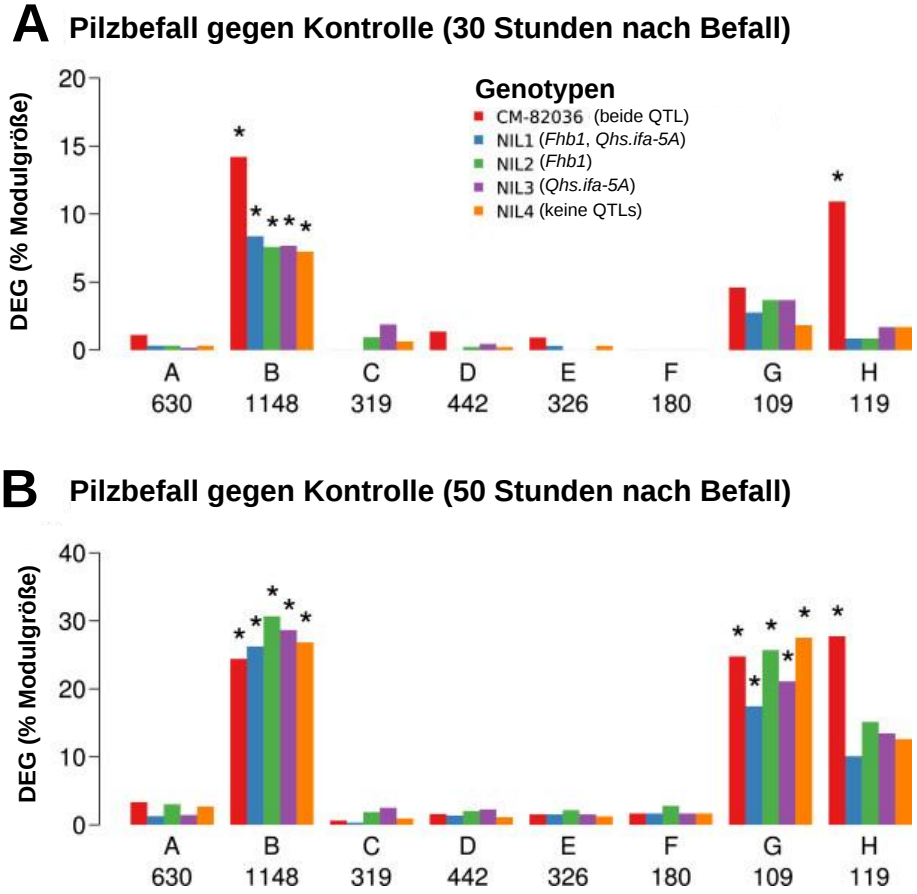

Abbildung 3.17: **Module des Ko-expressionsnetzwerks auf Basis der Expression von Weizengenen nach Pilzbefall**. Acht Module wurden bestimmt, die zwischen 109 und 1,148 Gene umfassten. Modul B und Modul G wiesen eine relativ zu Modulgröße hohe Anzahl an differentiell exprimierten Genen zwischen *Fg*-inokulierten und mit Wasser behandelten Pflanzen. (A) stellt die Anzahl an differentiell exprimierten Genen (DEG) pro Modul für den Zeitpunkt 30 Stunden nach Befall dar. (B) zum Zeitpunkt 50 Stunden nach Befall. Die Abbildung ist adaptiert nach Kugler et al. [2013] entnommen.

Weizenlinie	Cufflinks Gene	Gersten BBH	Gersten DEG	DEG in %
CM-82036	183,540	15,360	1,956	13%
NIL1	189,486	15,797	1,781	11%
NIL2	196,078	15,734	2,067	13%
NIL3	192,584	15,676	1,954	12%
NIL4	186,755	15,680	2,005	13%

Tabelle 3.17: **Exprimierte Gene in Weizen und im Vergleich zu Gerste**. Anzahl an Cufflinks und Gerstengene mit einem besten bidirektionalen Treffer zu einem Cufflinks Gen, sowie der Anteil an differentiell exprimierten Genen (DEG). Adaptiert nach Kugler et al. [2013]. NIL = nahezu isogene Linien, Cufflinks = Programm zur Bestimmung von Genmodellen durch RNA-seq Daten (Trapnell et al. [2012]).

Resistenz- oder Verteidigungsgenen beteiligt. Eine signifikante Anzahl an Gene wurde zwischen der Negativkontrolle (Wasser) gegenüber Pilzbefall differentiell exprimiert (Abbildung 3.18 und Kapitel 2.6). Ein deutlicher Anstieg von differentiell exprimierten Genen erfolgte von 30 Stunden nach 50 Stunden. Für Modul G (109 Gene) konnten keine signifikanten und überrepräsentierten Begriffen der Gene-Ontology-Datenbank bestimmt werden.

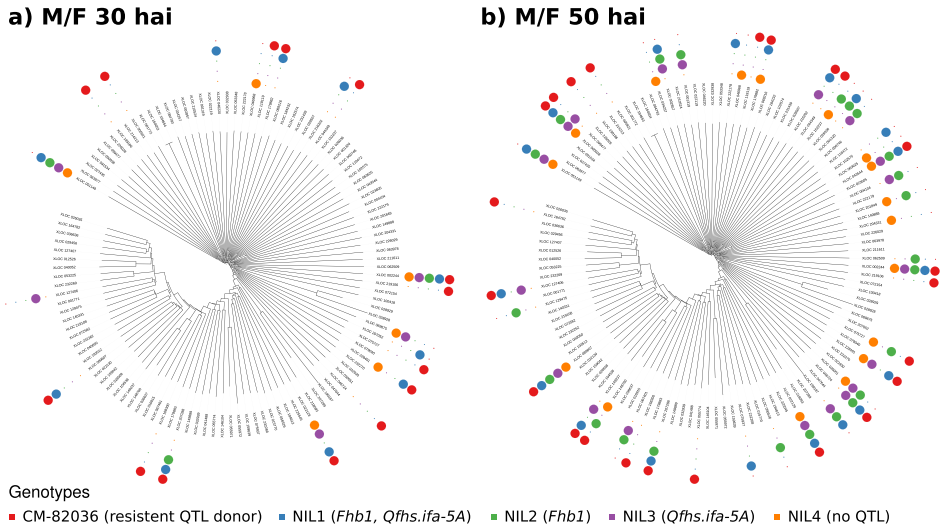

Abbildung 3.18: **Regulation von WRKY-Genen**. WRKY-Gene differentiell exprimiert 30 und 50 Stunden nach Befall. Entnommen aus Kugler et al. [2013]. Gene wurden in der Abbildung farblich markiert, wenn sie im Pilz/Wasser-Kontrast differentiell exprimiert wurden.

Die Verankerung von 15,719 Genen in Gerste ermöglichte es, die Vertei-

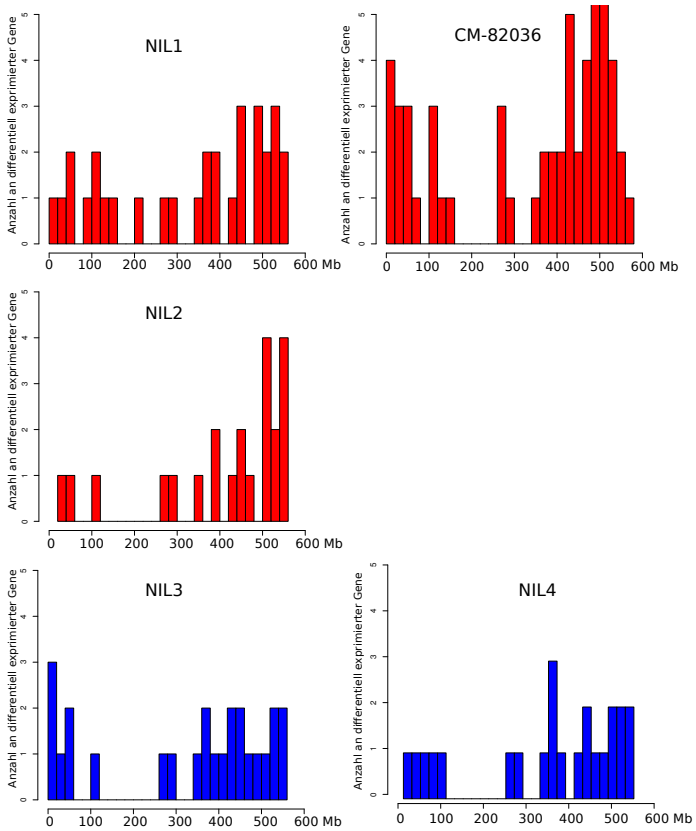

Abbildung 3.19: **Differentiell exprimierte Gene nach Pilzbefall**. Differentiell exprimierte Gene im Kontrast Pilzbefall und Negativkontrolle (Wasser) mit Sequenzhomologie zu Gerste wurden auf dem Gerste *gene-ome* verankert (Chromosom 3). Mit Rot dargestellte Histogramme stellen jene Linien mit *Fhb1* dar, blau jene Linien ohne *Fhb1*.

lung der differentiell exprimierten Gene einzugrenzen (Abbildung 3.19). Die starke Syntenie beider Genome (*Triticum aestivum* und *Hordeum vulgare*) sollte idealerweise zu einer Ansammlung der differentiell exprimierten Gene auf dem kurzen Arm von Gerstenchromosom 3H (Abbildung 3.19) führen und den *Fhb1*-QTL genetisch eingrenzen. Es ist zu erwarten, dass nicht nur das entscheidende Gen, sondern ein größerer genomischer Bereich in der Integressionslinien vorhanden ist. Die Darstellung aller fünf Linien zeigte allerdings keinen entscheidenden Unterschied in der Verteilung der differentiell exprimierten Gene zwischen den resistenten und empfänglichen Genotypen auf Gerstenchromosom 3H (Abbildung 3.19).

Das *IWGSC* stellte für alle 21 Chromosomenarme NGS Assemblies bereit und nutzte diese Contigs, um 124,201 Weizengene zu bestimmen (IWG-

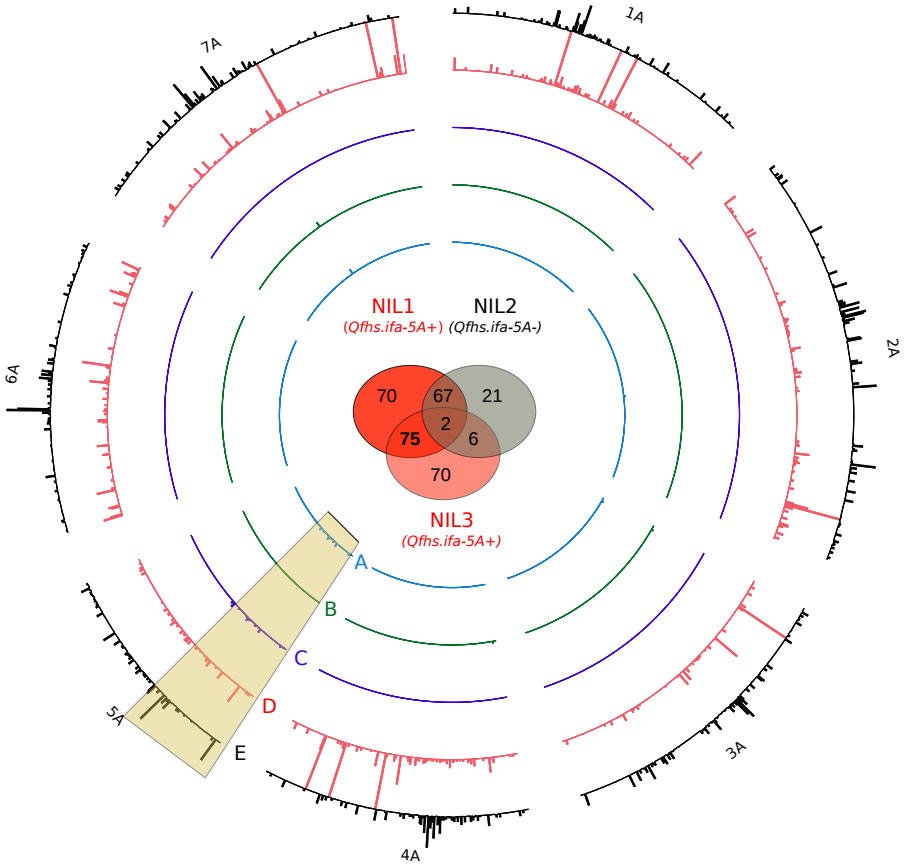

Abbildung 3.20: **Genetische Position differentiell exprimierter Gene auf dem Weizen A-Subgenom**. (A) differentiell exprimierte Gene zwischen NIL1/NIL4, (B) NIL2/NIL4, (C) NIL3/NIL4 und (D) CM-82036/NIL4. (A) - (C) weisen einen maximalen Wert von 20 Genen auf, (D) einen maximalen Wert von 100 Genen. (E) stellt die Genverteilung dar (0-600 Gene).

SC [2014]). Für die genetische Eingrenzung der beiden Resistenzbereiche wurden deshalb die 1.8 Tb RNA-seq Daten auf die in Chromosomenarme sortierten NGS contigs des *IWGSC* übertragen und die Expression jedes Weizengens bestimmt (siehe Methode 2.6). Anschließend wurden mit Hilfe von EdgeR (Nikolayeva and Robinson [2014]) differentiell exprimierte Gene bestimmt (durchgeführt durch Christian Ametz (*BOKU*)). Die Verteilung der differentiell exprimierten Gene von NIL1, NIL2, NIL3 sowie des Elter "CM-82036" gegenüber NIL4, der Weizenlinie, die keinen der beiden Resistenz-QTL aufweist, ist für A- und B-Subgenom aufgetragen (Abbildung 3.20, Abbildung 3.21). Der Wert pro genetischer Einheit (cM) wird durch die Anzahl der differentiell exprimierten Gene im Ver-

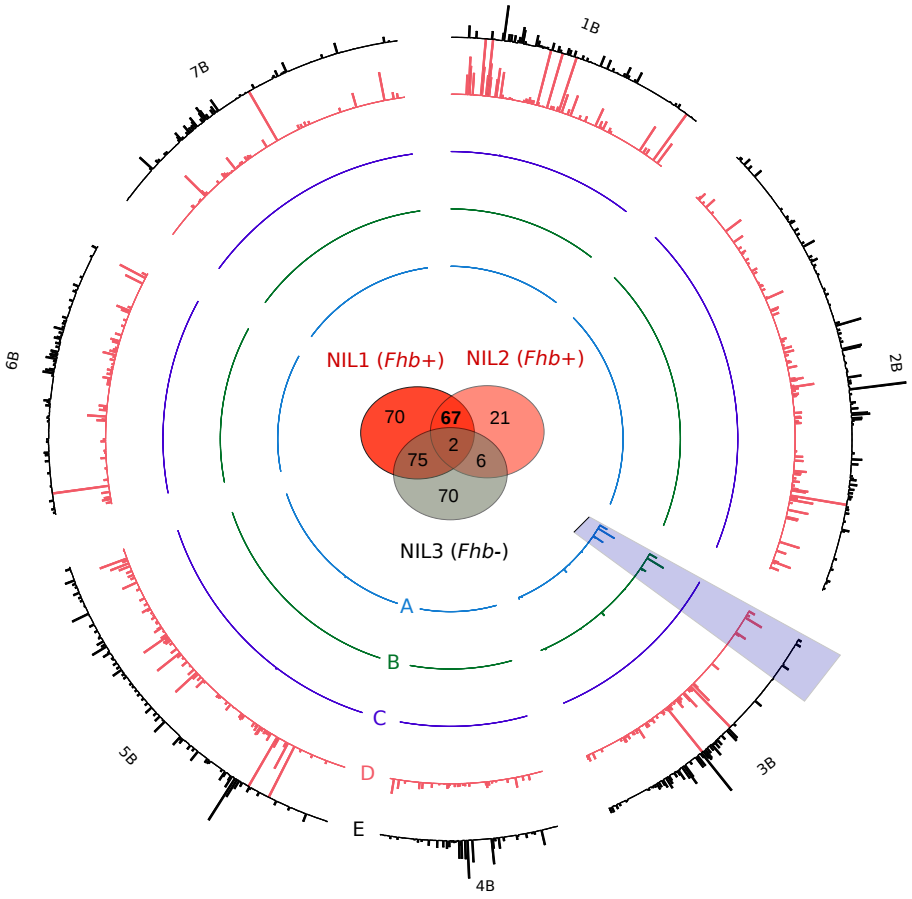

Abbildung 3.21: **Genetische Position differentiell exprimierter Gene auf dem Weizen B-Subgenom**. (A) differentiell exprimierte Gene zwischen NIL1/NIL4, (B) NIL2/NIL4, (C) NIL3/NIL4 und (D) CM-82036/NIL4. (A) - (C) weisen einen maximalen Wert von 20 Genen auf, (D) einen maximalen Wert von 100 Genen. (E) stellt die Genverteilung dar (0-600 Gene).

gleich zur Gesamtanzahl aller Gene in diesem Bereich bestimmt. Im QTL *Qfhs.ifa-5A* sind 75 (71%) aller differentiell exprimierten Gene zwischen NIL3(nur *Qfhs-ifa.5A*)/NIL4 gemeinsam mit Genen aus NIL1/NIL4. Im QTL *Fhb1* sind es 67 (70%) aller differentiell exprimierten Genen zwischen NIL2(nur *Fhb1*)/NIL4 gemeinsam mit Genen aus NIL1(beide QTL)/NIL4. Während für den QTL *Fhb1* eine klare genetische Eingrenzung auf dem kurzen Arm von 3BS möglich ist, deckt der auf Weizenchromosom 5A lokalisierte QTL einen größeren genetischen Bereich ab. Für das D-Subgenom konnten hingegen keine QTL-spezifischen Ansammlungen von differentiell exprimierten Genen festgestellt werden. Die eingefärbten Segmente aus den beiden Abbildungen (Abbildung 3.20, Abbildung 3.21) stellen die vermeint-

lichen Resistenzbereiche dar und sind deshalb mögliche Kandidaten für eine Funktion der QTL-spezifischen Resistenzantwort. Im QTL *Qfhs.ifa-5A* weisen nur wenige Gene wie Traes_5AL_9C9E55228 (Cellulose-Synthase Homolog), Traes_5AS_81202BD88 (Proteinkinase) und Traes_5AL_192FDBBCB (Funktion unbekannt) eine steigende Expression nach Pilzinfektion im Vergleich zur Negativkontrolle auf. Der Großteil der Gene ist exprimiert in den Linien mit dem QTL und nicht exprimiert in den Linien ohne dem QTL. Die Expression der Gene aus dem Vergleich der Linien mit beiden QTLs *Fhb1, Qfhs.ifa-5A* gegen Linien ohne diesen QTL wurde anschließend mit Hilfe des *RNASeqExpressionbrowsers* dargestellt (Abbildung 3.23 und Abbildung 3.22). Von 75 differentiell exprimierten Genen für den QTL

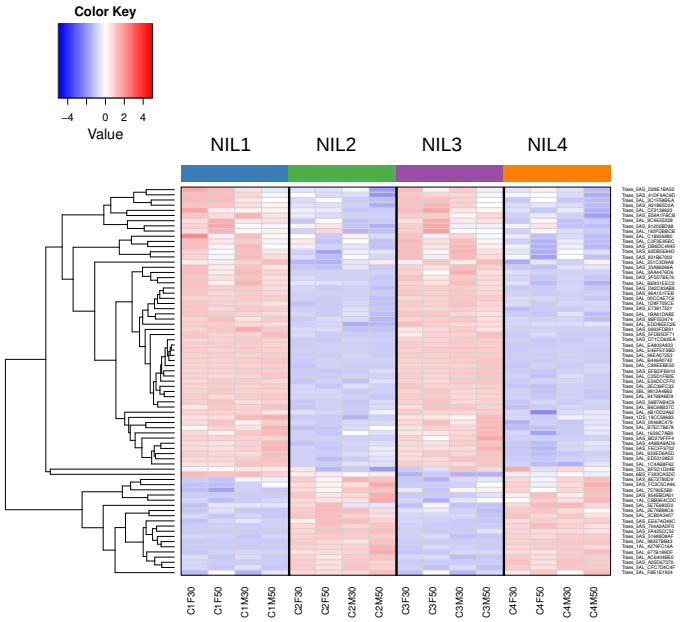

Abbildung 3.22: **Genkandidaten für den QTL *Qfhs.ifa-5A***. Differentiell exprimierte Gene aus dem Vergleich der Linien mit dem QTL *Fhb1* gegen Linien ohne diesen QTL und anschließend wurde die Expression der Gene des Chromosomes 5A mit Hilfe des *RNASeqExpressionbrowsers* dargestellt . In Blau dargestellt sind jene Gene, die eine deutlich niedrigere Expression im Vergleich zu durchschnittlichen Expression über alle Bedingungen aufweisen. In Rot sind jene Bedingungen pro Gen dargestellt, mit einer deutlich höheren Expression.

Qfhs.ifa-5A waren 46 Gene genetisch verankert, für Chromosom 3B wiesen 30 Gene (aus 67 Genen insgesamt) eine genetische Position auf. Die Gene Traes_3B_1396EA938 (DXP Reduktoisomerase) und Traes_3B_630DD22BF (Gen mit F-box Domäne) waren in den Linien mit dem *Fhb1*-QTL 50 Stunden nach Pilzbefall, im Vergleich zu allen anderen Bedingungen höher ex-

primiert. Sie wiesen aber auch eine hohe Expression in Negativkontrollen auf, während in den Linie ohne QTL kaum Expression gemessen wurde. Für Traes_3B_592CE6F7F (Thiolase) ist der Gegenteil der Fall: Eine starke Expression lag in den Linien ohne den QTL vor, kaum Expression in Linien mit diesem QTL.

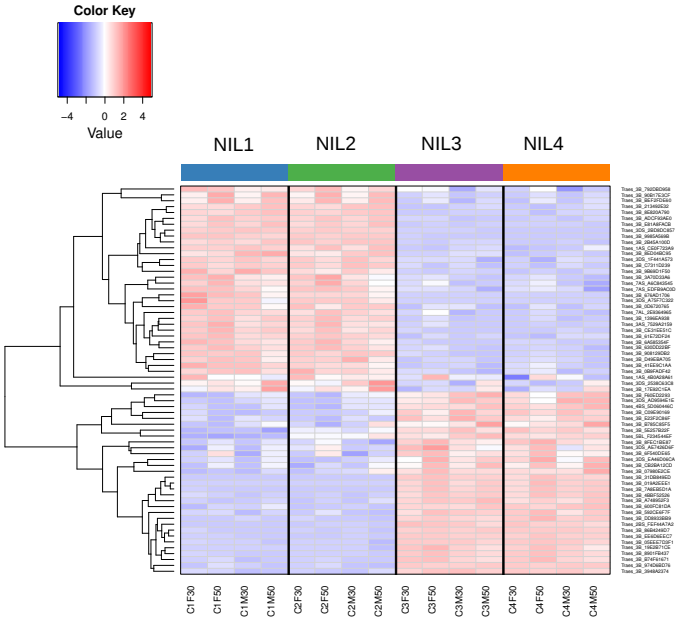

Abbildung 3.23: **Genkandidaten für den QTL *Fhb1***. Differentiell exprimierte Gene aus dem Vergleich der Linien mit dem QTL *Fhb1* gegen Linien ohne diesen QTL und anschließend wurde die Expressionswerte der Gene des Chromosomes 5A mit Hilfe des *RNASeqExpressionbrowsers* dargestellt. In Blau dargestellt sind jene Gene, die eine deutlich niedrigere Expression im Vergleich zu durchschnittlichen Expression über alle Bedingungen aufweisen. In Rot sind jene Bedingungen pro Gen dargestellt, mit einer deutlich höheren Expression.

Zusammenfassend kann gesagt werden, dass der in diesem Kapitel beschrittene Weg zu erfolgsversprechenden Resultaten und zu einer Detektion von möglichen Kandidatengenen für die Fusariosenresistenz führte. Dabei wurde das Gerste *gene-ome* wegen enger evolutionärer Verwandtschaft als Schablone für die mögliche Anordnung von Genen in Weizen (*T.aestivum*) verwendet und Expressionsdaten aus verschiedenen RIL-Linien mit und ohne Infektion auf dieses abgeleitete Weizengenommodel kartiert und Expressionsunterschiede untersucht. Erst mit den in Chromosomenarme sortieren *IWGSC* Genen konnten aber QTL-abhängige Gene bestimmt werden und diese genutzt werden, um die beiden QTLs genetisch einzugrenzen. Diese so eingegrenzten und aufgrund ihres Expressionsverhaltens eingegrenzten Kan-

didatengene können nun in weiteren Studien funktionell untersucht werden.

Kapitel 4

Diskussion

In der vorliegenden Arbeit wurde eine verankerte physikalische Karte für das Gerstegenom und für einzelne ausgewählte Weizenchromosomen bestimmt. Diese Ressourcen ermöglichen Regionen mit quantitativen Merkmalen (QTLs) zu lokalisieren und eine genetische Eingrenzung von Kandidatengenen. So konnten Resistenzbereiche gegen Mehltau (*Blumeria graminis*) lokalisiert werden und die molekularen Antworten in Weizen auf Ährenfusariosen (*Fusarium head blight*) untersucht werden. Im Folgenden wird die Bedeutung dieser Ressourcen diskutiert, zunächst für Gerste und anschließend für Weizen.

4.1 Bedeutung des Gerste *gene-ome* für Getreidearten

Die verankerte und mit Genen und NGS contigs versehene physikalische Karte in Gerste, das sogenannte Gerste *gene-ome*, ist die erste sehr umfangreiche genetische Anordnung eines größeren Getreidegenoms und der nächste Meilenstein in Gerste nach dem GenomeZipper (Mayer et al. [2011]). Das *gene-ome* umfasst eine Genom-ähnlich strukturierte Kollektion verschiedener Datenressourcen (*u. a.* genetische, genomische und Expressionsdaten) und erlaubte generelle Untersuchungen und Analysen, die überlicherweise ein vollständiges Genom als Basis voraussetzen. Mit dem *gene-ome* sind Vergleiche mit Wildgerste (*u. a. Hordeum vulgare* subsp. *spontaneum*) möglich, um Umstrukturierungen in Gerste sichtbar zu machen, die durch Züchtung und Züchtungsselektion entstanden sind (IBSC [2012]). Die unterschiedlichen Datensätze reichen von NGS-Assemblies von drei wirtschaftlich be-

deutenden Gerstenkultivaren (cv. “Morex”, cv. “Barke”, cv. “Bowman”), Expressionsdaten aus acht unterschiedlichen Wachstumsstadien bis hin zu einer großen Menge experimenteller und *in silico* Markern. Diese wurden aus publizierten Studien gesammelt oder im Zuge der Entwicklung und Verankerung der physikalischen Karte in Gerste neu erstellt. Das Gerste *gene-ome* erlaubte auch strukturelle Vergleiche gegen andere Getreidearten und Gerstensorten anzustellen (Luo et al. [2013], Martis et al. [2013]): In Roggen (Martis et al. [2013]) wurden die Pseudochromosomen aus Gerste und Markerkarten von unterschiedlichen Roggenkultivaren genutzt, um 17 syntenische Segmente in Roggen zu den Referenzgenomen aus Reis, *Sorghum* und *Brachypodium* zu bestimmen. Die unterschiedliche Konservierung der syntenischen Segmente, zusammen mit phylogenetischen Netzwerken basierend auf den orthologen Genen dieser Genome zeigte, dass moderner Roggen durch häufige Introgressionen aus unterschiedlichen Kultivaren geformt wurde. In Zeng et al. [2015] wurde das *gene-ome* genutzt, um frosttolerante Gerste aus dem Hochland in Tibet zu analysieren. Der Vergleich zu moderner Gerste erlaubte die Bestimmung von Genkandidaten für diese Toleranz.

4.1.1 Einfluss unterschiedlicher Klonbibliotheken

Für die Erstellung der FPC-Karte in Gerste wurden sieben unterschiedliche Klonbibliotheken mit einer durchschnittlich 14-fachen Genomabdeckung berücksichtigt. Die große Anzahl an Klonbibliotheken wurde gewählt, um sicherzustellen, dass alle genomischen Bereiche auch vollständig repräsentiert sind und nach Sequenzierung des minimalen überspannenden Pfads die FP contigs zu Chromosomen zusammengefasst werden. Die Klonbibliothek HVVMRXALLHA nutzt das Restriktionsenzym *Mbo*1 und deckt alleine 4.4 Gb der 4.99 Gb großen FPC-Karte ab, gefolgt von der HVVMRXALLEA-Bibliothek (3.9 Gb) (Restriktionsenzym *Eco*R1). Die restlichen Bibliotheken erreichen durch geringe Klonzahlen nur rund 50% der physikalischen Karte. Die Zahl an schimärischen FP contigs nach manueller Überprüfung fiel mit 171 contigs gering aus. Dies ist zum einen bedingt durch die hohe Genomabdeckung bei der Erstellung der physikalischen Karte und zum anderen durch die Verwendung von vier unterschiedlichen Restriktionsschnittstellen. Das Gerste *gene-ome* zeigt, dass für diploide Getreidegenome eine Klonassemblierung auch unabhängig von einer Trennung einzelner Chromosomen möglich ist. Im hexaploiden Weizen (*T. aestivum*) hingegen, erfolgte eine Trennung einzelner Chromosomen durch die Durchflusszytometrie,

um die sehr ähnlichen Subgenome voneinander zu trennen. Die Verwendung von sieben unterschiedlichen Klonbibliotheken führte zu längeren FP contig-Längen, weil durch unterschiedliche Bandenmuster und durch eine höhere Genomabdeckung eine größere Anzahl an signifikanten Klonüberlappungen bestimmt werden konnte. Zwei Klonbibliotheken, die beispielsweise durch *Eco*R1 und *Mbo*1 behandelt werden, sollten allerdings ausreichen, um eine robuste physikalische Karte in Gerste zu erstellen. Die Verwendung lediglich einer Klonbibliothek führt zu FP contigs mit geringeren Längen und erhöht das Risiko schimärischer Contigs, wie Gerste festgestellt wurde. In *Aegilops tauschii* wurden später ebenfalls sieben Klonbibliotheken genutzt und in Summe 461,706 Klone zu FP contigs zusammengefasst, im Vergleich zu den 517,202 Klonen in Gerste und ebenfalls vier unterschiedliche Restriktionsenzyme genutzt. In *Aegilops tauschii* wurde mit einem Sulstonscore von 10^{-22} eine weniger stringente Assemblierung als in Gerste (10^{-45}) gewählt, wodurch nur 3,153 FP contigs anstatt 9,265 FP contigs in Gerste assembliert wurden. In dieser Arbeit durchgeführte Analysen (Kapitel 3.1.10) zeigten, dass in der Gerste eine Assemblierung von Klonen in FP contigs vermutlich auch mit weniger stringenten Kriterien möglich wäre.

4.1.2 Markerdatensätze - Einfluss auf die Verankerung

Zunächst wurden genetische Karten zur Bestimmung und Suche nach bestimmten QTLs erstellt (*u. a.* Laurie et al. [1995]). In den letzten Jahren wurden Markerkarten mit jeweils einigen tausenden Markersequenzen publiziert (*u. a.* Stein et al. [2007], Potokina et al. [2008], Sato et al. [2009]). Besonders hoch-auflösende Markerkarten wurden in Close et al. [2009] und Comadran et al. [2012] bereitgestellt. In Close et al. [2009] wurde eine Konsensuskarte für Gerste durch Integration mehrerer genetischer Karten erstellt. Diese Konsensuskarte bündelt ≈22,000 SNPs auf Basis von Gersten EST-Sequenzen und 4,596 Amplikons und basiert auf europäischen und US-Sorten. Diese Konsensuskarte stellte 2,943 SNP Sequenzen mit einer genetischen Markerlänge von 1,099 cM bereit und erlaubte syntenische Vergleiche zu Reis. Diese Karte wurde später als Referenzkarte für den Gersten Genomzipper (Mayer et al. [2011]) verwendet. Für die Integration der genetischen Karten in dieser Studie wurde die Markerkarte mit höchster Auflösung (SM6, Comadran et al. [2012]) als Referenzkarte genutzt, und weitere experimentelle und *in silico* Markerkarten wurden integriert. Diese Karte basierte auf der Kreuzung Morex x Bar-

ke. Kultivar Morex wurde ebenfalls für die BAC-Bibliothek verwendet. In dieser Arbeit wurde gezeigt, dass genische Marker nur einen geringen Beitrag zur Verankerung von FP contigs im Zentromer-nahen Bereiche leisten. *GBS*-basierte Markerkarten (SM7-SM10) erlauben auch FP contigs ohne sequenzierte genische Bereiche zu verankern und ermöglichten die genetische Verankerung von 3.9 Gb der physikalischen Karte. Die Markerzuordnungen der Referenzkarte sowie aller anderen experimentellen und *in silico* Marker wurden in einem zentralen Datenarchiv zusammengefasst. Unter `ftp://ftpmips.helmholtz-muenchen.de/plants/barley/public_data/anchoring/in_silico_marker` liegen die Markertreffer zu FP contigs, NGS contigs sowie zu Genen bereit.

4.1.3 Von ESTs hin zur Genannotation

In HarvEST (`http://harvest.ucr.edu/`) sind EST-Datensätze aus zehn unterschiedlichen Pflanzengenomen gebündelt, unter anderem auch für Gerste und Weizen. Dieser EST-Datensatz wurde zunächst für die Erstellung von Markerkarten und Microarrays genutzt. Im Jahr 2011 wurden 24,783 Gersten Volllängen-cDNA Sequenzen bestimmt und schließlich durch das *gene-ome* 26,159 Gene mit definierten Exon-Intron Strukturen. Diese Gene wurden auf Basis von acht Wachstumsstadien für die stadienspezifische Expression analysiert. 15,719 Gene wurden im Rahmen dieser Arbeit physikalisch verankert, ein Großteil davon wies einen direkten Bezug zu einem FP contig auf. Durch Syntenie zu Reis, *Brachypodium* und *Sorghum* wurde die Zahl auf 20,411 Gene gesteigert. Durch Bereitstellung von genomischen Sequenzen können bereits teilweise Promoteranalysen (Dey et al. [2014]) durchgeführt und Kandidaten-Gene genetisch gekoppelt werden. Ressourcen wie BARLEYMAP (Cantalapiedra et al. [2015]) und *chromoWIZ* (Nussbaumer et al. [2014b]) erlauben die Suche nach diesen Genen.

4.1.4 Verhältnis zwischen genetischer und physikalischer Karte

Die verankerte physikalische Karte zeigt, dass in ≈40% des Genoms eine genetische Reihung von FP contigs unmöglich ist. Im Vergleich zum gesamten Genom liegt im wenig rekombinierenden zentromernahen Bereich eine $20x$ geringere Rekombinationsrate vor (Baker et al. [2014]). Für bestimmte Regionen entspricht 1 cM nur wenigen hundert kb, in anderen Bereichen hingegen genomischen Bereichen von 200 Mb (IBSC [2012]). Hoch-auflösen-

de Karten erlauben es, diese Bereiche zu schmälern und helfen, zu isolierende Zielgene wie beispielsweise Resistenzgene genetisch einzugrenzen. Sie erfordern allerdings eine hohe Populationsgröße, um auch seltene Rekombinationsereignisse zu messen. Durch die physikalischen Karte zusammen mit einer hochauflösenden genetischen Karte wird die Suche nach Genen auf bestimmte Klone reduziert. Zusätzliche Marker können generiert werden, wenn die physikalische Distanz noch zu groß ist. Ein Beispiel, wie Ressourcen des *IBSC* genutzt wurden, um einzelne QTL genetisch einzugrenzen, wurde in Silvar et al. [2013] beschrieben: In dieser Studie wurden drei QTLs spanischer Gerstenlinien gegen "Echten Gerstenmehltau" (*Blumeria graminis f. sp. hordei*) untersucht. Die drei QTLs befinden sich auf 7HS, 7HL und 6HL und konnten mit Hilfe des GenomeZippers und der physikalischen Karte in Gerste genauer untersucht werden. Zusätzlich wurden FP contigs bestimmt, die mit den drei QTLs überlappen. Anschließend wurden alle Gene und NGS contigs innerhalb des genetischen Intervalls bestimmt. Für 7HS, 7HL und 6HL konnte der genomische Bereich auf 4.0, 3.7 sowie 3.2 Mb eingegrenzt werden. Zusätzlich wurden 21, 10 bzw. 16 Kandidatengene bestimmt. Diese Gene können in anschließenden Studien genauer untersucht werden.

4.2 Einfluss der verankerten Weizenchromosomen

Die Arbeiten an Weizenchromosom 6A (Poursarebani et al. [2014]) sowie die Klonassemblierungen der Weizenchromosomen 1D, 4D und 6D (Sehgal et al. [In Vorbereitung]) erlaubten es, neue Methodiken aus Gerste wie der Integration von experimentellen und *in silico* Markerkarten, *CarmA* und die NGS-Anreicherung der physikalischen Karte zu nützen und den Einfluss von neu entwickelten Verfahren wie die WGP^{TM}-Technologie und 3D-Klonpools zu untersuchen. In Weizenchromosom 6A wurde die "Whole genome profiling" (WGP^{TM})-Technologie genutzt. In diesem Verfahren werden die umliegenden Bereiche eines Restriktionsenzyms sequenziert ("Sequenz-*Tags*"). Diese 100 bp langen Sequenzen werden durch deutlich längere Sequenzen erweitert und stellen für jeden Klon im Durchschnitt 25 dieser Sequenz-*Tags* bereit. Durch zu geringe Qualität der letzten 50 bp der WGP^{TM}-*Tags* konnten aber nur 50 bp zur Integration von Daten genutzt werden. Zu diesem Zeitpunkt lagen noch keine längeren Sequenzen für Weizenchromosomen vor. Um diese Lücken zu schließen, wurden die nächsten verwandten Genome aus *Triticum urartu* und *Aegilops tauschii* sowie Daten des *IWGSC* auf physikalische FP contigs übertragen und ≈662 Mb aus 1,048 Mb der kumulativen Länge der

FP contig-Karte dieses Chromosoms genetisch verankert. Syntenische Vergleiche zu Gerste erhöhten die Anzahl an genetisch verankerten physikalischen Contigs auf 831 Mb. Der Anteil an verankerten FP contigs mit 78% der kumulativen Länge der FPC-Karte ist vergleichbar mit den Weizenchromosomen aus 1BL (74%), 1AS (74%) und 1AL (75%). Weizenchromosom 6A zeigt, dass die Erstellung einer physikalischen Karte für einen einzelnen Chromosomenarm in Weizen auch mit Hilfe nur einer Klonbibliothek bei ≈8x Genomabdeckung möglich ist, während in Gerste und *Aegilops tauschii* für ein diploides Genom sieben Klonbibliotheken und vier Restriktionsenzyme genutzt wurden. Die Verwendung von zwei Assemblierungsprogrammen erlaubte außerdem, die bestmögliche FP contig-Karte zu konstruieren. LTC (Frenkel et al. [2010]) erwies sich gegenüber FPC (Soderlund et al. [1997]) als besonders robust. Diese Feststellung resultierte aus der geringeren Anzahl an schimärischen Contigs und vor allen aufgrund einer größeren Anzahl an integrierten Klonen. Ein Grund der Überlegenheit von LTC gegenüber FPC liegt an den hoch-abundanten Bandenmuster, die in hoch-repetitiven Genomen wie Gerste, Weizen und Mais vorliegen (Frenkel et al. [2010]). FPC hat Schwierigkeiten, die Signifikanz der Überlappungen einzuschätzen, hingegen erlaubt LTC durch das Verfahren der topologischen Strukturbestimmung von Contigs auch für diese Banden eine bessere Einschätzung der Signifikanz. Zusätzlich bietet LTC grafische Darstellungen, um potentielle problematische Klone, sogenannte "Q-Klone" zu bestimmen, und FP contigs gegenfalls zu trennen. Durch das Fehlen von längeren genomischen Bereichen in Weizen, zeigte Chromosom 6A, dass Ressourcen aus den Vorläufergenomen mit deutlichen größeren NGS contigs genutzt werden können, um die physikalische Karte in Weizen maximal auszuschöpfen.

In Weizen 1D, 4D und 6D wurden 7,064 Klone des MTP bei geringer Genomabdeckung sequenziert und im Rahmen dieser Arbeit assembliert. Hierzu wurden Sequenzen aus einzelnen 3D-Klonpools sequenziert, und in dieser Arbeit assembliert und einem Klon zugeordnet; ein Vorgang der als Klondekonvolution bezeichnet wird. Durch das *IWGSC* wurden Chromosomenarm sortierte Contigs bereitgestellt. Diese Daten erlaubten die Trennung der drei Chromosomen 1D, 4D und 6D. Im Zuge der Konstruktion der BAC-Bibliotheken konnten diese drei Chromosomen aufgrund des ähnlichen DNA-Gehalt nicht voneinander getrennt (Janda et al. [2004]). Die Ressourcen aus dem *IWGSC* erlaubten schließlich durch *CarmA* eine Trennung der physikalischen FP contigs in die drei Chromosomen. Die Ergebnisse aus den Chromosomen 1D, 4D und 6D zeigen, dass der Großteil (80.2%) der Sequenzen zu

einem bestimmten Klon zugewiesen wird. Durch Zuordnung der Sequenzen zu einzelnen Klonen, der Anwendung von *CarmA* und die Berücksichtigung des nahe verwandten Genoms aus *Aegilops tauschii* ermöglichte aber eine *in silico* Trennung der einzelnen Chromosomen.

4.2.1 Resistenz gegen Ährenfuriosen

In dieser Arbeit wurde die Resistenz von unterschiedlichen Weizenkultivaren gegen Ährenfuriosen beschrieben (Kugler et al. [2013]). Diese Studie stellte eine der ersten Beschreibungen unter Zuhilfenahme von RNA-seq Daten aus unterschiedlich resistenten Pflanzen dar. Bisherige Studien wurden aufgrund des Fehlens einer Referenzsequenz größtenteils auf Basis von Microarrays durchgeführt (*u. a.* Schweiger et al. [2013]). In der ersten von zwei in dieser Arbeit beschriebenen Studien (Kugler et al. [2013]) wurden Gene auf die zu diesem Zeitpunkt aktuellste Genomreferenz von Weizen bestimmt. Diese Referenz repräsentierte ein Assembly aus Sequenzen der 454-Sequenzierung mit $5x$-facher Genomabdeckung. Dieser Datensatz erlaubte aber noch keine umfangreiche Genannotation. Es lag an der geringen Genomabdeckung und an den annähernd identischen Subgenomen des allohexaploiden Weizen und führte zu kurzen Contig-Längen. Um proteinkodierende Gene und funktionelle Annotationen zu nutzen wurden deshalb 233k potentielle Weizengene auf dem Assembly der 454-Sequenzierung den 26,159 Gerstengenen zugewiesen. Wegen der kurzen evolutionären Distanz der beiden Getreiden von rund 12 Millionen Jahren (Wicker et al. [2009a]), kann erwartet werden, dass der Großteil der Gene konserviert ist. Die rund 16k Gene mit einem Gerstenortholog wurden anschließend für ein Ko-expressionsnetzwerk verwendet, wodurch die genomweite Expression der Genotypen untersucht wurde, die sich in einem oder in beiden Resistenzbereichen unterscheiden. Zwei der acht Module zeigten unterschiedliche Expressionswerte bei Pilzbefall. Zu diesem Zeitpunkt konnten aber noch keine Unterschiede festgestellt werden, die die Eigenschaft einer der Resistenzbereiche auf den Chromosomen 3B und 5A erklären würden. In Kugler et al. [In Vorbereitung] wurden die Chromosomenarm sortierten Contigs genutzt und die Expressionsdaten aus Kugler et al. [2013] neu analysiert. Meine Analysen aus Kugler et al. [In Vorbereitung] zeigten eine Anhäufung von differentiell exprimierten Genen in den zu erwarteten Chromosomen 3B und 5A (vgl. Abbildung 3.20 und Abbildung 3.21). Durch funktionelle Analysen auf Basis von 1.8 Tb RNA-Seq Daten und durch die in Chromosomen getrennte Datensätze konnten Kandidaten-Gene

auf den beschriebenen chromosomalen Bereichen der beiden QTLs bestimmt und genetisch eingrenzt werden. Auf diese Studien aufbauenden Analysen unter Zuhilfenahme von mehreren Zeitreihen sollten dadurch einen besseren Einblick auf die Funktion der beiden QTLs bieten.
Die Referenzsequenzen basieren allerdings auf dem infektionsanfälligen Kultivar "Chinese Spring", während die resistenten Linien vom Weizenkultivar "Sumai-3" abgeleitet sind. Deshalb muss die Interpretation der Ergebnisse mit Vorsicht behandelt werden. Es ist unklar, ob beide Genotypen auch tatsächlich eine perfekte Kollinearität in diesem genomischen Bereich aufweisen und ob das entscheidende Gen sogar spezifisch für diesen Genotyp ist.

4.3 Herausforderungen und Limitierungen

In dieser Arbeit konnte ich durch die Integration von unterschiedlichen heterogenen Datenressourcen eine genomische Referenz für Gerste und für ausgewählte Weizenchromosomen (1D, 4D, 6A, 6D) erstellen. Die genetisch verankerten physikalischen Karten in Verbindung mit der Auswahl des minimalen zusammengehörenden Pfads sind der letzte notwendige Schritt zu den chromosomalen Sequenzen. Die Bestimmung der Chromosomen aus Gerste und Weizen kann aufgrund ihres hohen Anteils an repetitiven Elementen und damit verbundenen Kosten in der Sequenzierung und Schwierigkeiten in der Assemblierung nur im Rahmen von größeren Konsortien (Gersten- und Weizengenomsequenzierungsprojekt) durchgeführt werden. Für Gerste und Weizen, im Gegensatz zu deutlich kleineren Genomen, reichen Hochdurchsatz-Sequenzierungsmethoden zur Bestimmung des Chromosoms trotz hoher Genomabdeckung nicht aus. Für diese größeren Pflanzengenome ist eine verankerte physikalische Karte zur Bestimmung der chromosomalen Sequenz eine Notwendigkeit, um die Komplexität in der Assemblierung von genomischer Sequenz zu reduzieren. 2,670,738 NGS contigs, die 1,869 Mb des Genoms repräsentieren, stehen nur mehr 9,265 physikalischen Contigs gegenüber. Durch eine physikalische Karte muss die Assemblierung des Genoms nicht mehr auf Basis eines 5 Gb großen Genoms, sondern innerhalb von im Durschnitt 538 kb großen DNA-Fragmenten durchgeführt werden. Bei ausreichender Genomabdeckung erlaubt es auch, hoch-repetitive Bereiche zu überspannen, beispielsweise durch Sequenzen aus *PacBio*-Sequenzierung. Die Qualität und Vollständigkeit des minimalen überspannenden Pfads stellt eine besondere Bedeutung dar, weil dieser als Vorlage zur Sequenzierung des

gesamten Genoms dient.
In dieser Arbeit wurden Syntenievergleiche zu kleineren und nahe verwandten Referenzgenomen durchgeführt. Der Vergleich wies keine genomischen Lücken auf; ein Indiz dafür, dass die physikalische Karte das Gerstengenom gut abdeckt wird (IBSC [2012]). In dieser Arbeit wurden auch neu entwickelte Verfahren wie die Chromosomenarm-Zuordnung (*CarmA*) präsentiert und überprüft, ob konsistente Chromosomenarm-Zuordnungen zwischen dem physikalischen Contig zu seinen verankerten Sequenzen wie Klonenden und Genen bestehen. Die Datensätze aus dem *IWGSC* wurden direkt auf die physikalische Karte übertragen, während in Gerste das vollständige Potential aller Datensätze nicht genutzt werden konnte. Die genetische Verankerung von NGS contigs wurde nur durchgeführt, wenn eine Chromosomenarm-Zuordnung vorlag. So konnten nur ≈350 Mb der 1.8 Gb Contigs aus Gerstenkultivar "Morex" genutzt werden. Diese Anzahl an genetisch verankerten Sequenzen wurde in einer späteren Studie deutlich erhöht, die Anzahl an physikalischen Contigs konnte allerdings mit dieser Strategie nicht erhöht werden (Mascher et al. [2013], Ariyadasa et al. [2014]). Die in dieser Arbeit verbliebenen physikalischen Contigs ohne genetischer Marker wiesen meist nur wenige oder keine sequenzierten Bereiche auf und umfassen annähernd 1,000 FP contigs (IBSC [2012]). In einigen Fällen konnten allerdings experimentelle Marker einen FP contig auch ohne sequenzierte Bereiche verankern, wenn ein Marker durch Hybridisierung auf einem Klon zugewiesen wurde. Genische Marker haben Limitierungen in der Anzahl an verankerten Sequenzen und erlaubten im Zuge der ersten Verankerungen, noch ohne *GBS* Marker, maximal 2 Gb der physikalischen Karte genetisch zu positionieren, im Vergleich zu 3.9 Gb in Verbindung mit *GBS* Markerkarten. *GBS* Marker verankern auch viele physikalische Contigs exklusiv (IBSC [2012]). Die *GBS*-Technologie und ihr großer Nutzen zur Verankerung großer und komplexer Pflanzengenome wurde auch in Mais beschrieben (Elshire et al. [2011]), dem bisher größten sequenzierten und assemblierten Genom. Durch die Erfolge in der Verankerung von Gerste wurden *GBS*-basierte Ansätze auch in einigen weiteren Genomen wie Hafer und Sojabohne angewandt (*u.a.* Huang et al. [2014], Elmer et al. [2015]).
Nach Fertigstellung dieser Arbeit erfolgten große Fortschritte innerhalb der genetischen Verankerung von Gersten NGS contigs: 1.2 Gb des 1.8 Gb Assembly wurde genetisch verankert (Mascher et al. [2013]). In dieser Arbeit fiel der Grad an Verankerungen deutlich geringer aus, weil eine eindeutige Chromosomenarm-Zuordnung für einen FP contigs, NGS contig oder Gen

gefordert wurde. Durch Bereitstellung der mit Sequenzen angereicherten FP contigs konnte ich aber bereits eine Ressource schaffen, die es erlaubt, Kandidatensequenzen einem bestimmten FP contigs zugewiesen. So können Klone gezielt bestimmt werden, wie in Silvar et al. [2013] dargestellt wurde. Dort wurden drei Resistenzbereiche in spanischen Landrassen auf spezifische genomische Bereiche eingegrenzt. Diese Bereiche und überlappende Klone können, sofern sie nicht bereits sequenziert wurden, aus den Klonbibliotheken gezielt gewählt und sequenziert werden. Außerdem können mit den Ergebnissen dieser Arbeit Transkriptomanalysen auf der genomischen Referenzsequenz beider Genome durchgeführt werden und Kandidatengene bestimmt sowie gegebenenfalls funktionell überprüft werden (*u. a.* Dey et al. [2014]).

Die Zielsetzung war, die beiden Genome (Gerste, Weizen) systematisch zugänglich zu machen und führte zur Entwicklung des Webtools *chromoWIZ* (Nussbaumer et al. [2014b]), mit dem durch funktionelle und Sequenzbasierte Methoden genetische/genomische Bereiche markiert werden, in denen eine besonders hohe Anzahl an Kandidatengenen vorliegt. Die Erstellung eines weiteren Webtools, dem *RNASeqExpressionBrowser* (Nussbaumer et al. [2014a]) erlaubte es außerdem, Expressionsprofile einzelner Gene zu suchen.

Durch die Bereitstellung der Genomassemblierungsdatensätze und der Annotation von Genen auf dem Gerstenkultivar Morex konnte in dieser Arbeit gezeigt werden, dass Gene in beiden Genomen an den proximalen und distalen Chromosomenenden lokalisiert sind und in ihrer Dichte angereichert sind und es zu einem deutlichen Abfall der Gendichte an den Perizentromerischen Bereichen kommt und damit den typischen Verlauf von Gendichten in den Gräsern zeigte. Die verankerte physikalische Karte in Gerste ist dabei hilfreich, um die Umstrukturierungen in Weizen, beispielsweise aus Weizenchromosom 4A (Hernandez et al. [2012]) genauer genetisch und genomisch einzugrenzen und durch die besonders hohe Anzahl an genetischen Gerstenmarkern weitere Weizensequenzen zu verankern. Ein Beispiel dafür, wie Gerste hilft, um Weizenverankerungen zu erhöhen, wurde in dieser Arbeit am Weizenchromosom 6A (Kapitel 3.2) gezeigt. Die große Übereinstimmung der genomischen Ressourcen aus Gerste mit den Weizenmarkern erlaubte es, den Ansatz zu validieren und bietet eine zusätzliche Möglichkeit bei der Suche nach Kandidatengenen.

4.4 Zusammenfassung und Ausblick

Die verankerten physikalischen Karten in Gerste und Weizen weisen Limitierungen auf, die durch fragmentarische Genomassemblierungsdatensätze bedingt sind. FP contigs können nur zusammengefasst werden, wenn Klone vollständig sequenziert vorliegen. NGS-Sequenzierungen basierend auf einem genomweiten Datensatz führen zu kollabierten Assemblierungen der hochrepetitiven Bereiche, erlauben aber eine annähernd komplette Genannotation, weil Gene zumeist in den weniger hoch-repetitiven Bereichen vorliegen. Im Gegensatz zu NGS contigs ermöglicht aber nur eine vollständige genomische Sequenz die eindeutige Lokalisierung von Genen und Eingrenzung von Kandidatengenen. Mit der Verfügbarkeit eines sequenzierten und assemblierten Referenzgenoms können interessante Bereiche aber genetisch eingegrenzt werden. Trotz dieser Eingrenzung verbleiben aber in rekombinationsarmen Regionen mehrere hundert Gene als potentielle Kandidaten.
In Feuillet et al. [2011] wurde gezeigt, dass seit Bereitstellung der genomischen Sequenzen in Reis, bis 2007 18 QTLs kloniert wurden. Bis ins Jahr 2000 konnten hingegen nur zwei QTLs bestimmt werden. Für Weizen, das ebenfalls angeführt wurde, konnte bis ins Jahr 2007 erst ein QTL kloniert werden.
Aus diesem Grund soll das Genom zukünftig vollständig bestimmt werden, um die Lokalisierung von weiteren QTLs kosten- und zeiteffizient durchzuführen. Mit Hilfe von unterschiedlichen hochdurchsatzbasierten Datensätzen mit variierenden Abständen zwischen Einzelsequenzen, neuen Technologien (*GBS*, Introgressionslinien) sowie unterschiedlichen Sequenzierungstechnologien ist abzusehen, dass die Assemblierung eines gesamten Genoms erfolgreich durchgeführt werden kann. Bestimmte FP contigs im zentromernahen Bereich müssen allerdings durch Hybridisierungen (*u. a.* Fluoreszenz-in-situ-Hybridisierung (FISH) (Karafiátová et al. [2013])) oder Optical Map Ansätze verankert werden, um Limitierungen bedingt durch geringe genetische Rekombination adäquat zu kompensieren.
Die Analyse zur Pilzinfektion in Weizen, beschrieben in zwei Studien (Kugler et al. [2013], Kugler et al. [In Vorbereitung]) stellte ein Anwendungsbeispiel für die Ergebnisse der verankerten physikalischen Karten dar. Methoden ((*chromoWIZ* (Nussbaumer et al. [2014b]), *RNASeqExpressionsBrowser* (Nussbaumer et al. [2014a]))), die in dieser Arbeit erstellt wurden, erlaubten die genetische Eingrenzung sowie funktionelle Analyse der beiden Resistenzbereiche *Qfhs.ifa-5A* und *Fhb1*, die hauptverantwortlich für Resistenz gegen

den Pilz *Fusarium graminearum* und gegen die Verbreitung des von diesem Pilz erzeugten Toxins (DON) sind. Zunächst wurden Ressourcen aus Gerste genutzt, einem Genom, das im Gegensatz zu Weizen ein deutlich kleineres und diploides Genom aufweist. Die Gerstenressourcen erlaubten aus vielen tausenden Genkandidaten in Weizen, die auf einem Assembly mit sehr kleinen Contigs annotiert wurden, proteinkodierende Gene zu bestimmen und damit zu diesem Zeitpunkt die aktuellste und umfangreichste Verankerung und funktionelle Annotation eines diploiden Getreidengenoms. Zu diesem Zeitpunkt lagen in Weizen noch keine größeren genomischen Sequenzen vor. Später, mit Verfügbarkeit der mit Genen annotierten und in Chromosomenarme sortierten Contigs in Weizen, durch das *IWGSC*, wurden schließlich diese Bereiche genetisch eingegrenzt, wodurch 67 Kandidaten Gene für *Fhb1* und 75 Kandidaten Gene für *Qfhs.ifa-5A* bestimmt werden konnten. Diese Gene können nun, in aktuelen Studien und unter Zuhilfenahme von weiteren Zeitpunkten (3h, 6h, 12h, 24h, 36h) analysiert und experimentell getestet werden.

Literaturverzeichnis

AGI. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796–815, Dec 2000.

A. Alexeyenko, B. Nystedt, F. Vezzi, E. Sherwood, R. Ye, B. Knudsen, M. Simonsen, B. Turner, P. de Jong, C. C. Wu, and J. Lundeberg. Efficient de novo assembly of large and complex genomes by massively parallel sequencing of Fosmid pools. *BMC Genomics*, 15:439, 2014.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, Oct 1990.

R. Ariyadasa, M. Mascher, T. Nussbaumer, D. Schulte, Z. Frenkel, N. Poursarebani, R. Zhou, B. Steuernagel, H. Gundlach, S. Taudien, M. Felder, M. Platzer, A. Himmelbach, T. Schmutzer, P. E. Hedley, G. J. Muehlbauer, U. Scholz, A. Korol, K. F. Mayer, R. Waugh, P. Langridge, A. Graner, and N. Stein. A sequence-ready physical map of barley anchored genetically by two million single-nucleotide polymorphisms. *Plant Physiol.*, 164 (1):412–423, Jan 2014.

A. Badr, K. Muller, R. Schafer-Pregl, H. El Rabey, S. Effgen, H. H. Ibrahim, C. Pozzi, W. Rohde, and F. Salamini. On the origin and domestication history of Barley (Hordeum vulgare). *Mol. Biol. Evol.*, 17(4):499–510, Apr 2000.

K. Baker, M. Bayer, N. Cook, S. Dreissig, T. Dhillon, J. Russell, P. E. Hedley, J. Morris, L. Ramsay, I. Colas, R. Waugh, B. Steffenson, I. Milne, G. Stephen, D. Marshall, and A. J. Flavell. The low-recombining pericentromeric region of barley restricts gene diversity and evolution but not gene expression. *Plant J.*, 79(6):981–992, Sep 2014.

I. Birol, A. Raymond, S. D. Jackman, S. Pleasance, R. Coope, G. A. Taylor, M. M. Yuen, C. I. Keeling, D. Brand, B. P. Vandervalk, H. Kirk,

P. Pandoh, R. A. Moore, Y. Zhao, A. J. Mungall, B. Jaquish, A. Yanchuk, C. Ritland, B. Boyle, J. Bousquet, K. Ritland, J. Mackay, J. Bohlmann, and S. J. Jones. Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. *Bioinformatics*, 29(12): 1492–1497, Jun 2013.

S. Bolot, M. Abrouk, U. Masood-Quraishi, N. Stein, J. Messing, C. Feuillet, and J. Salse. The 'inner circle' of the cereal genomes. *Curr. Opin. Plant Biol.*, 12(2):119–125, Apr 2009.

R. Brenchley, M. Spannagl, M. Pfeifer, G. L. Barker, R. D'Amore, A. M. Allen, N. McKenzie, M. Kramer, A. Kerhornou, D. Bolser, S. Kay, D. Waite, M. Trick, I. Bancroft, Y. Gu, N. Huo, M. C. Luo, S. Sehgal, B. Gill, S. Kianian, O. Anderson, P. Kersey, J. Dvorak, W. R. McCombie, A. Hall, K. F. Mayer, K. J. Edwards, M. W. Bevan, and N. Hall. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, 491(7426):705–710, Nov 2012.

C. Cantalapiedra, R. Boudiar, A. Casas, E. Igartua, and B. Contreras-Moreira. Barleymap: physical and genetic mapping of nucleotide sequences and annotation of surrounding loci in barley. *Molecular Breeding*, 35(1):13, 2015. ISSN 1380-3743. doi: 10.1007/s11032-015-0253-1. URL `http://dx.doi.org/10.1007/s11032-015-0253-1`.

J. Chen, Q. Huang, D. Gao, J. Wang, Y. Lang, T. Liu, B. Li, Z. Bai, J. Luis Goicoechea, C. Liang, C. Chen, W. Zhang, S. Sun, Y. Liao, X. Zhang, L. Yang, C. Song, M. Wang, J. Shi, G. Liu, J. Liu, H. Zhou, W. Zhou, Q. Yu, N. An, Y. Chen, Q. Cai, B. Wang, B. Liu, J. Min, Y. Huang, H. Wu, Z. Li, Y. Zhang, Y. Yin, W. Song, J. Jiang, S. A. Jackson, R. A. Wing, J. Wang, and M. Chen. Whole-genome sequencing of Oryza brachyantha reveals mechanisms underlying Oryza genome evolution. *Nat Commun*, 4:1595, 2013.

X. Chen, C. A. Hackett, R. E. Niks, P. E. Hedley, C. Booth, A. Druka, T. C. Marcel, A. Vels, M. Bayer, I. Milne, J. Morris, L. Ramsay, D. Marshall, L. Cardle, and R. Waugh. An eQTL analysis of partial resistance to Puccinia hordei in barley. *PLoS ONE*, 5(1):e8598, 2010.

F. Choulet, T. Wicker, C. Rustenholz, E. Paux, J. Salse, P. Leroy, S. Schlub, M. C. Le Paslier, G. Magdelenat, C. Gonthier, A. Couloux, H. Budak, J. Breen, M. Pumphrey, S. Liu, X. Kong, J. Jia, M. Gut, D. Brunel, J. A.

Anderson, B. S. Gill, R. Appels, B. Keller, and C. Feuillet. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell*, 22(6):1686–1701, Jun 2010.

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

T. J. Close, P. R. Bhat, S. Lonardi, Y. Wu, N. Rostoks, L. Ramsay, A. Druka, N. Stein, J. T. Svensson, S. Wanamaker, S. Bozdag, M. L. Roose, M. J. Moscou, S. Chao, R. K. Varshney, P. Szucs, K. Sato, P. M. Hayes, D. E. Matthews, A. Kleinhofs, G. J. Muehlbauer, J. DeYoung, D. F. Marshall, K. Madishetty, R. D. Fenton, P. Condamine, A. Graner, and R. Waugh. Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics*, 10:582, 2009.

CoGePedia. Sequenced plant genomes @ONLINE, 2014. URL `http://genomevolution.org/wiki/index.php/Sequenced_plant_genomes`.

J. Comadran, B. Kilian, J. Russell, L. Ramsay, N. Stein, M. Ganal, P. Shaw, M. Bayer, W. Thomas, D. Marshall, P. Hedley, A. Tondelli, N. Pecchioni, E. Francia, V. Korzun, A. Walther, and R. Waugh. Natural variation in a homolog of Antirrhinum CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. *Nat. Genet.*, 44(12):1388–1392, Dec 2012.

P. A. Cuthbert, D. J. Somers, J. Thomas, S. Cloutier, and A. Brulé-Babel. Fine mapping Fhb1, a major gene controlling fusarium head blight resistance in bread wheat (Triticum aestivum L.). *Theor. Appl. Genet.*, 112 (8):1465–1472, May 2006.

K. Devos, G. Moore, and M. Gale. Conservation of marker synteny during evolution. *Euphytica*, 85:367–372, 1995.

S. Dey, M. Wenig, G. Langen, S. Sharma, K. G. Kugler, C. Knappe, B. Hause, M. Bichlmeier, V. Babaeizad, J. Imani, I. Janzik, T. Stempfl, R. Huckelhoven, K. H. Kogel, K. F. Mayer, and A. C. Vlot. Bacteria-Triggered

Systemic Immunity in Barley Is Associated with WRKY and ETHYLENE RESPONSIVE FACTORs But Not with Salicylic Acid. *Plant Physiol.*, 166(4):2133–2151, Dec 2014.

S. Diguistini, N. Y. Liao, D. Platt, G. Robertson, M. Seidel, S. K. Chan, T. R. Docking, I. Birol, R. A. Holt, M. Hirst, E. Mardis, M. A. Marra, R. C. Hamelin, J. Bohlmann, C. Breuil, and S. J. Jones. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biol.*, 10(9):R94, 2009.

Y. Ding, M. D. Johnson, W. Q. Chen, D. Wong, Y. J. Chen, S. C. Benson, J. Y. Lam, Y. M. Kim, and H. Shizuya. Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. *Genomics*, 74(2):142–154, Jun 2001.

A. N. Egan, J. Schlueter, and D. M. Spooner. Applications of next-generation sequencing in plant biology. *Am. J. Bot.*, 99(2):175–185, Feb 2012.

M. El Baidouri and O. Panaud. Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biol Evol*, 5(5):954–965, 2013.

I. Elmer, S. Humira, and B. Francois. Association mapping of QTLs for sclerotinia stem Rot resistance in a collection of soybean plant introductions using a genotyping by sequencing (GBS) approach. *BMC Plant Biol.*, 15 (1):5, Jan 2015.

R. J. Elshire, J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, 6(5):e19379, 2011.

FAOSTAT. Faostat @ONLINE, 2013. URL `http://faostat.fao.org`.

C. Feuillet, J. E. Leach, J. Rogers, P. S. Schnable, and K. Eversole. Crop genome sequencing: lessons and rationales. *Trends Plant Sci.*, 16(2):77–88, Feb 2011.

Z. Frenkel, E. Paux, D. Mester, C. Feuillet, and A. Korol. LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes. *BMC Bioinformatics*, 11:584, 2010.

S. A. Goff et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science*, 296(5565):92–100, Apr 2002.

P. Hernandez, M. Martis, G. Dorado, M. Pfeifer, S. Gálvez, S. Schaaf, N. Jouve, H. Simková, M. Valárik, J. Dolezel, and K. F. Mayer. Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J.*, 69(3):377–386, Feb 2012.

A. Hossain, J. A. Teixeira da Silva, M. V. Lozovskaya, and V. P. Zvolinsky. High temperature combined with drought affect rainfed spring wheat and barley in South-Eastern Russia: I. Phenology and growth. *Saudi J Biol Sci*, 19(4):473–487, Oct 2012.

Y. F. Huang, J. A. Poland, C. P. Wight, E. W. Jackson, and N. A. Tinker. Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. *PLoS ONE*, 9(7):e102448, 2014.

E. Ibarra-Laclette, E. Lyons, G. Hernández-Guzmán, C. A. Pérez-Torres, L. Carretero-Paulet, T. H. Chang, T. Lan, A. J. Welch, M. J. Juárez, J. Simpson, A. Fernández-Cortés, M. Arteaga-Vázquez, E. Góngora-Castillo, G. Acevedo-Hernández, S. C. Schuster, H. Himmelbauer, A. E. Minoche, S. Xu, M. Lynch, A. Oropeza-Aburto, S. A. Cervantes-Pérez, M. de Jesús Ortega-Estrada, J. I. Cervantes-Luevano, T. P. Michael, T. Mockler, D. Bryant, A. Herrera-Estrella, V. A. Albert, and L. Herrera-Estrella. Architecture and evolution of a minute plant genome. *Nature*, 498(7452):94–98, Jun 2013.

IBSC. A physical, genetic and functional sequence assembly of the barley genome. *Nature*, 491(7426):711–716, Nov 2012.

IRGSP. The map-based sequence of the rice genome. *Nature*, 436(7052): 793–800, Aug 2005.

IWGSC. A chromosome-based draft sequence of the hexaploid bread wheat (triticum aestivum) genome. *Science*, 345(6194), July 2014.

J. Janda, J. Bartos, J. Safár, M. Kubaláková, M. Valárik, J. Číhalíková, H. Simková, M. Caboche, P. Sourdille, M. Bernard, B. Chalhoub, and J. Dolezel. Construction of a subgenomic BAC library specific for chromosomes 1D, 4D and 6D of hexaploid wheat. *Theor. Appl. Genet.*, 109 (7):1337–1345, Nov 2004.

J. Jia, S. Zhao, X. Kong, Y. Li, G. Zhao, W. He, R. Appels, M. Pfeifer, Y. Tao, X. Zhang, R. Jing, C. Zhang, Y. Ma, L. Gao, C. Gao, M. Spannagl, K. F. Mayer, D. Li, S. Pan, F. Zheng, Q. Hu, X. Xia, J. Li, Q. Liang, J. Chen, T. Wicker, C. Gou, H. Kuang, G. He, Y. Luo, B. Keller, Q. Xia, P. Lu, J. Wang, H. Zou, R. Zhang, J. Xu, J. Gao, C. Middleton, Z. Quan, G. Liu, J. Wang, H. Yang, X. Liu, Z. He, L. Mao, J. Wang, C. Feuillet, K. Eversole, B. Keller, J. Dvorak, B. Gill, Y. Ogihara, and R. Appels. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, 496(7443):91–95, Apr 2013.

M. Karafiátová, J. Bartoš, D. Kopecký, L. Ma, K. Sato, A. Houben, N. Stein, and J. Doležel. Mapping nonrecombining regions in barley using multicolor FISH. *Chromosome Res.*, Sep 2013.

P. Karlovsky. Biological detoxification of the mycotoxin deoxynivalenol and its use in genetically engineered crops and feed additives. *Appl. Microbiol. Biotechnol.*, 91(3):491–504, Aug 2011.

K. G. Kugler, G. Siegwart, T. Nussbaumer, C. Ametz, M. Spannagl, B. Steiner, M. Lemmens, K. F. Mayer, H. Buerstmayr, and W. Schweiger. 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*, 14:728, 2013.

K. G. Kugler, T. Nussbaumer, B. Warth, S. Sharma, C. Ametz, A. Simader, A. Parich, M. Lemmens, R. Schuhmacher, R. Krska, H. Buerstmayr, K. Mayer, and W. Schweiger. Multilayered dissection of the molecular bread wheat (Triticum aestivum) response to Fusarium graminearum. In Vorbereitung.

P. Langfelder and S. Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, Apr 2012.

D. A. Laurie, N. Pratchett, J. W. Snape, and J. H. Bezant. RFLP mapping of five major genes and eight quantitative trait loci controlling flowering time in a winter x spring barley (Hordeum vulgare L.) cross. *Genome*, 38 (3):575–585, Jun 1995.

I. Letunic and P. Bork. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.*, 39(Web Server issue):W475–478, Jul 2011.

H. Q. Ling, S. Zhao, D. Liu, J. Wang, H. Sun, C. Zhang, H. Fan, D. Li, L. Dong, Y. Tao, C. Gao, H. Wu, Y. Li, Y. Cui, X. Guo, S. Zheng, B. Wang, K. Yu, Q. Liang, W. Yang, X. Lou, J. Chen, M. Feng, J. Jian, X. Zhang, G. Luo, Y. Jiang, J. Liu, Z. Wang, Y. Sha, B. Zhang, H. Wu, D. Tang, Q. Shen, P. Xue, S. Zou, X. Wang, X. Liu, F. Wang, Y. Yang, X. An, Z. Dong, K. Zhang, X. Zhang, M. C. Luo, J. Dvorak, Y. Tong, J. Wang, H. Yang, Z. Li, D. Wang, A. Zhang, and J. Wang. Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature*, 496(7443):87–90, Apr 2013.

L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.*, 2012:251364, 2012.

S. Lonardi, D. Duma, M. Alpert, F. Cordero, M. Beccuti, P. R. Bhat, Y. Wu, G. Ciardo, B. Alsaihati, Y. Ma, S. Wanamaker, J. Resnik, S. Bozdag, M. C. Luo, and T. J. Close. Combinatorial pooling enables selective sequencing of the barley gene space. *PLoS Comput. Biol.*, 9(4):e1003010, Apr 2013.

Q. Long, F. A. Rabanal, D. Meng, C. D. Huber, A. Farlow, A. Platzer, Q. Zhang, B. J. Vilhjálmsson, A. Korte, V. Nizhynska, V. Voronin, P. Korte, L. Sedman, T. Mandáková, M. A. Lysak, U. Seren, I. Hellmann, and M. Nordborg. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat. Genet.*, 45(8):884–890, Aug 2013.

M. C. Luo, Y. Q. Gu, F. M. You, K. R. Deal, Y. Ma, Y. Hu, N. Huo, Y. Wang, J. Wang, S. Chen, C. M. Jorgensen, Y. Zhang, P. E. McGuire, S. Pasternak, J. C. Stein, D. Ware, M. Kramer, W. R. McCombie, S. F. Kianian, M. M. Martis, K. F. Mayer, S. K. Sehgal, W. Li, B. S. Gill, M. W. Bevan, H. Simková, J. Doležel, S. Weining, G. R. Lazo, O. D. Anderson, and J. Dvorak. 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.*, 110(19):7940–7945, May 2013.

R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu,

C. Han, D. W. Cheung, S. M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T. W. Lam, and J. Wang. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1):18, 2012.

K. Madishetty, P. Condamine, J. T. Svensson, E. Rodriguez, and T. J. Close. An improved method to identify BAC clones using pooled overgos. *Nucleic Acids Res.*, 35(1):e5, 2007.

T. Marcussen, S. R. Sandve, L. Heier, M. Spannagl, M. Pfeifer, K. S. Jakobsen, B. B. Wulff, B. Steuernagel, K. F. Mayer, O. A. Olsen, K. F. Mayer, J. Rogers, J. Dolezel, C. Pozniak, K. Eversole, C. Feuillet, B. Gill, B. Friebe, A. J. Lukaszewski, P. Sourdille, T. R. Endo, M. Kubalakova, J. ?ihalikova, Z. Dubska, J. Vrana, R. Sperkova, H. Simkova, M. Febrer, L. Clissold, K. McLay, K. Singh, P. Chhuneja, N. K. Singh, J. Khurana, E. Akhunov, F. Choulet, A. Alberti, V. Barbe, P. Wincker, H. Kanamori, F. Kobayashi, T. Itoh, T. Matsumoto, H. Sakai, T. Tanaka, J. Wu, Y. Ogihara, H. Handa, P. Maclachlan, A. Sharpe, D. Klassen, D. Edwards, J. Batley, O. A. Olsen, S. R. Sandve, S. Lien, B. Steuernagel, B. Wulff, M. Caccamo, S. Ayling, R. H. Ramirez-Gonzalez, B. J. Clavijo, J. Wright, M. Pfeifer, M. Spannagl, M. M. Martis, M. Mascher, J. Chapman, J. A. Poland, U. Scholz, K. Barry, R. Waugh, D. S. Rokhsar, G. J. Muehlbauer, N. Stein, H. Gundlach, M. Zytnicki, V. Jamilloux, H. Quesneville, T. Wicker, P. Faccioli, M. Colaiacovo, A. M. Stanca, H. Budak, L. Cattivelli, N. Glover, L. Pingault, E. Paux, S. Sharma, R. Appels, M. Bellgard, B. Chapman, T. Nussbaumer, K. C. Bader, H. Rimbert, S. Wang, R. Knox, A. Kilian, M. Alaux, F. Alfama, L. Couderc, N. Guilhot, C. Viseux, M. Loaec, B. Keller, and S. Praud. Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, 345(6194):1250092, Jul 2014.

M. M. Martis, R. Zhou, G. Haseneyer, T. Schmutzer, J. Vrana, M. Kubalakova, S. Konig, K. G. Kugler, U. Scholz, B. Hackauf, V. Korzun, C. C. Schon, J. Dolezel, E. Bauer, K. F. Mayer, and N. Stein. Reticulate evolution of the rye genome. *Plant Cell*, 25(10):3685–3698, Oct 2013.

M. Mascher, G. J. Muehlbauer, D. S. Rokhsar, J. Chapman, J. Schmutz, K. Barry, M. Muñoz-Amatriaín, T. J. Close, R. P. Wise, A. H. Schulman, A. Himmelbach, K. F. Mayer, U. Scholz, J. A. Poland, N. Stein, and

R. Waugh. Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J.*, Sep 2013.

T. Matsumoto, T. Tanaka, H. Sakai, N. Amano, H. Kanamori, K. Kurita, A. Kikuta, K. Kamiya, M. Yamamoto, H. Ikawa, N. Fujii, K. Hori, T. Itoh, and K. Sato. Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol.*, 156(1):20–28, May 2011.

K. F. Mayer, S. Taudien, M. Martis, H. Simková, P. Suchánková, H. Gundlach, T. Wicker, A. Petzold, M. Felder, B. Steuernagel, U. Scholz, A. Graner, M. Platzer, J. Dolezel, and N. Stein. Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol.*, 151(2):496–505, Oct 2009.

K. F. Mayer, M. Martis, P. E. Hedley, H. Simková, H. Liu, J. A. Morris, B. Steuernagel, S. Taudien, S. Roessner, H. Gundlach, M. Kubaláková, P. Suchánková, F. Murat, M. Felder, T. Nussbaumer, A. Graner, J. Salse, T. Endo, H. Sakai, T. Tanaka, T. Itoh, K. Sato, M. Platzer, T. Matsumoto, U. Scholz, J. Dolezel, R. Waugh, and N. Stein. Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell*, 23(4): 1249–1263, Apr 2011.

B. C. Meyers, S. Scalabrin, and M. Morgante. Mapping and sequencing complex genomes: let's get physical! *Nat. Rev. Genet.*, 5(8):578–588, Aug 2004.

M. J. Moscou, N. Lauter, B. Steffenson, and R. P. Wise. Quantitative and qualitative stem rust resistance factors in barley are associated with transcriptional suppression of defense regulons. *PLoS Genet.*, 7(7):e1002208, Jul 2011.

O. Nikolayeva and M. D. Robinson. edgeR for Differential RNA-seq and ChIP-seq Analysis: An Application to Stem Cell Biology. *Methods Mol. Biol.*, 1150:45–79, 2014.

T. Nussbaumer, M. M. Martis, S. K. Roessner, M. Pfeifer, K. C. Bader, S. Sharma, H. Gundlach, and M. Spannagl. MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res.*, 41 (Database issue):D1144–1151, Jan 2013.

T. Nussbaumer, K. G. Kugler, K. C. Bader, S. Sharma, M. Seidel, and

K. F. Mayer. RNASeqExpressionBrowser - A web interface to browse and visualize high-throughput expression data. *Bioinformatics*, May 2014a.

T. Nussbaumer, K. G. Kugler, W. Schweiger, K. C. Bader, H. Gundlach, M. Spannagl, N. Poursarebani, M. Pfeifer, and K. Mayer. chromoWIZ: a web tool to query and visualize chromosome-anchored genes from cereal and model genomes. *BMC Plant Biol.*, 14(1):348, Dec 2014b.

E. Paux, P. Sourdille, J. Salse, C. Saintenac, F. Choulet, P. Leroy, A. Korol, M. Michalak, S. Kianian, W. Spielmeyer, E. Lagudah, D. Somers, A. Kilian, M. Alaux, S. Vautrin, H. Bergès, K. Eversole, R. Appels, J. Safar, H. Simkova, J. Dolezel, M. Bernard, and C. Feuillet. A physical map of the 1-gigabase bread wheat chromosome 3B. *Science*, 322(5898):101–104, Oct 2008.

M. Pfeifer, M. Martis, T. Asp, K. F. Mayer, T. Lübberstedt, S. Byrne, U. Frei, and B. Studer. The perennial ryegrass GenomeZipper: targeted use of genome resources for comparative grass genomics. *Plant Physiol.*, 161(2):571–582, Feb 2013.

R. Philippe, F. Choulet, E. Paux, J. van Oeveren, J. Tang, A. H. Wittenberg, A. Janssen, M. J. van Eijk, K. Stormo, A. Alberti, P. Wincker, E. Akhunov, E. van der Vossen, and C. Feuillet. Whole Genome Profiling provides a robust framework for physical mapping and sequencing in the highly complex and repetitive wheat genome. *BMC Genomics*, 13:47, 2012.

J. A. Poland, P. J. Brown, M. E. Sorrells, and J. L. Jannink. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE*, 7(2):e32253, 2012.

B. Poppenberger, F. Berthiller, D. Lucyshyn, T. Sieberer, R. Schuhmacher, R. Krska, K. Kuchler, J. Glossl, C. Luschnig, and G. Adam. Detoxification of the Fusarium mycotoxin deoxynivalenol by a UDP-glucosyltransferase from Arabidopsis thaliana. *J. Biol. Chem.*, 278(48):47905–47914, Nov 2003.

E. Potokina, A. Druka, Z. Luo, R. Wise, R. Waugh, and M. Kearsey. Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *Plant J.*, 53 (1):90–101, Jan 2008.

N. Poursarebani, T. Nussbaumer, H. Simkova, J. Safár, H. Witsenboer, J. van Oeveren, J. Doležel, K. F. Mayer, N. Stein, and T. Schnurbusch. Whole-genome profiling and shotgun sequencing delivers an anchored, gene-decorated, physical map assembly of bread wheat chromosome 6A. *Plant J.*, 79(2):334–347, Jul 2014.

V. Prade. Eine Validierung von CarmA: Read based Chromosome arm Assignment. 2013.

M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13:341, 2012.

C. Saintenac, D. Jiang, S. Wang, and E. Akhunov. Sequence-based mapping of the polyploid wheat genome. *G3 (Bethesda)*, 3(7):1105–1114, Jul 2013.

K. Sato, N. Nankaku, and K. Takeda. A high-density transcript linkage map of barley derived from a single population. *Heredity (Edinb)*, 103(2): 110–117, Aug 2009.

K. Sato, T. J. Close, P. Bhat, M. Muñoz-Amatriaín, and G. J. Muehlbauer. Single nucleotide polymorphism mapping and alignment of recombinant chromosome substitution lines in barley. *Plant Cell Physiol.*, 52(5):728–737, May 2011a.

K. Sato, Y. Motoi, N. Yamaji, and H. Yoshida. 454 sequencing of pooled BAC clones on chromosome 3H of barley. *BMC Genomics*, 12:246, 2011b.

S. Sato et al. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400):635–641, May 2012.

P. S. Schnable et al. The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–1115, Nov 2009.

D. Schulte, T. J. Close, A. Graner, P. Langridge, T. Matsumoto, G. Muehlbauer, K. Sato, A. H. Schulman, R. Waugh, R. P. Wise, and N. Stein. The international barley sequencing consortium–at the threshold of efficient access to the barley genome. *Plant Physiol.*, 149(1):142–147, Jan 2009.

D. Schulte, R. Ariyadasa, B. Shi, D. Fleury, C. Saski, M. Atkins, P. deJong, C. C. Wu, A. Graner, P. Langridge, and N. Stein. BAC library resources

for map-based cloning and physical map construction in barley (Hordeum vulgare L.). *BMC Genomics*, 12:247, 2011.

W. Schweiger, B. Steiner, C. Ametz, G. Siegwart, G. Wiesenberger, F. Berthiller, M. Lemmens, H. Jia, G. Adam, G. J. Muehlbauer, D. P. Kreil, and H. Buerstmayr. Transcriptomic characterization of two major Fusarium resistance quantitative trait loci (QTLs), Fhb1 and Qfhs.ifa-5A, identifies novel candidate genes. *Mol. Plant Pathol.*, 14(8):772–785, Oct 2013.

S. Sehgal et al. The anchored physical maps of 1D, 4D and 6D. In Vorbereitung.

K. Shirasu, A. H. Schulman, T. Lahaye, and P. Schulze-Lefert. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.*, 10(7):908–915, Jul 2000.

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

C. Soderlund, I. Longden, and R. Mott. FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.*, 13(5):523–535, Oct 1997.

N. Stein, M. Prasad, U. Scholz, T. Thiel, H. Zhang, M. Wolf, R. Kota, R. K. Varshney, D. Perovic, I. Grosse, and A. Graner. A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. *Theor. Appl. Genet.*, 114(5):823–839, Mar 2007.

B. Steuernagel, S. Taudien, H. Gundlach, M. Seidel, R. Ariyadasa, D. Schulte, A. Petzold, M. Felder, A. Graner, U. Scholz, K. F. Mayer, M. Platzer, and N. Stein. De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. *BMC Genomics*, 10:547, 2009.

The International Brachypodium Initiative. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, 463(7282): 763–768, Feb 2010.

J. D. Thompson, T. J. Gibson, and D. G. Higgins. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*, Chapter 2:Unit 2.3, Aug 2002.

C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 7(3):562–578, Mar 2012.

C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, 31(1):46–53, Jan 2013.

J. van Oeveren, M. de Ruiter, T. Jesse, H. van der Poel, J. Tang, F. Yalcin, A. Janssen, H. Volpin, K. E. Stormo, R. Bogden, M. J. van Eijk, and M. Prins. Sequence-based physical mapping of complex genomes by whole genome profiling. *Genome Res.*, 21(4):618–625, Apr 2011.

N. Vitulo, A. Albiero, C. Forcato, D. Campagna, F. Dal Pero, P. Bagnaresi, M. Colaiacovo, P. Faccioli, A. Lamontanara, H. Šimková, M. Kubaláková, G. Perrotta, P. Facella, L. Lopez, M. Pietrella, G. Gianese, J. Doležel, G. Giuliano, L. Cattivelli, G. Valle, and A. M. Stanca. First survey of the wheat chromosome 5A composition through a next generation sequencing approach. *PLoS ONE*, 6(10):e26421, 2011.

J. Vrana, M. Kubalakova, H. Simkova, J. Cihalikova, M. A. Lysak, and J. Dolezel. Flow sorting of mitotic chromosomes in common wheat (Triticum aestivum L.). *Genetics*, 156(4):2033–2041, Dec 2000.

W. Wang, G. Haberer, H. Gundlach, C. Gläßer, T. Nussbaumer, M. C. Luo, A. Lomsadze, M. Borodovsky, R. A. Kerstetter, J. Shanklin, D. W. Byrant, T. C. Mockler, K. J. Appenroth, J. Grimwood, J. Jenkins, J. Chow, C. Choi, C. Adam, X. H. Cao, J. Fuchs, I. Schubert, D. Rokhsar, J. Schmutz, T. P. Michael, K. F. Mayer, and J. Messing. The Spirodela polyrhiza genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nat Commun*, 5:3311, Feb 2014.

S. N. Wegulo. Factors influencing deoxynivalenol accumulation in small grain cereals, 2012. URL `http://www.mdpi.com/2072-6651/4/11/1157`.

T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A. H. Schulman. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, 8(12):973–982, Dec 2007.

T. Wicker, S. G. Krattinger, E. S. Lagudah, T. Komatsuda, M. Pourkheirandish, T. Matsumoto, S. Cloutier, L. Reiser, H. Kanamori, K. Sato, D. Perovic, N. Stein, and B. Keller. Analysis of intraspecies diversity in wheat and barley genomes identifies breakpoints of ancient haplotypes and provides insight into the structure of diploid and hexaploid triticeae gene pools. *Plant Physiol.*, 149(1):258–270, Jan 2009a.

T. Wicker, S. Taudien, A. Houben, B. Keller, A. Graner, M. Platzer, and N. Stein. A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J.*, 59(5):712–722, Sep 2009b.

J. Yu et al. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science*, 296(5565):79–92, Apr 2002.

X. Zeng, H. Long, Z. Wang, S. Zhao, Y. Tang, Z. Huang, Y. Wang, Q. Xu, L. Mao, G. Deng, X. Yao, X. Li, L. Bai, H. Yuan, Z. Pan, R. Liu, X. Chen, Q. WangMu, M. Chen, L. Yu, J. Liang, D. DunZhu, Y. Zheng, S. Yu, Z. LuoBu, X. Guang, J. Li, C. Deng, W. Hu, C. Chen, X. TaBa, L. Gao, X. Lv, Y. B. Abu, X. Fang, E. Nevo, M. Yu, J. Wang, and N. Tashi. The draft genome of Tibetan hulless barley reveals adaptive patterns to the high stressful Tibetan Plateau. *Proc. Natl. Acad. Sci. U.S.A.*, 112(4): 1095–1100, Jan 2015.

W. C. Zhou, F. L. Kolb, G. H. Bai, L. L. Domier, and J. B. Yao. Effect of individual Sumai 3 chromosomes on resistance to scab spread within spikes and deoxynivalenol accumulation within kernels in wheat. *Hereditas*, 137 (2):81–89, 2002.