GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Veröffentlichung der Ergebnisse von Forschungsvorhaben im BMBF-Programm

B I O L O G I E

Forschungsvorhaben: GABI-FUTURE-Verbundvorhaben: „Erschließung des genetischen Potentials von Roggen mittels Etablierung einer Ressource zur funktionellen Genomanalyse des exprimierten (‘EXPRESSed‘) Anteils des Roggengenoms (GABI-RYE EXPRESS)“ (Teilprojekt A)

Förderkennzeichen: 0315063 A

Zuwendungsempfänger: Technische Universität München

Ausführende Stelle: Technische Universität München - Wissenschaftszentrum Weihenstephan - Forschungsdepartment für Pflanzenwissenschaften – Lehrstuhl für Pflanzenzüchtung, 85354 Freising

Projektleitung: Dr. Eva Bauer

Laufzeit: 01.01.2008 – 30.06.2011

„Das diesem Bericht zugrundeliegende BMBF-Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 0315063A gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor“.

Berichtsblatt

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

3. Titel

a) Forschungsvorhaben: GABI-FUTURE-Verbundvorhaben: „Erschließung des genetischen Potentials von Roggen mittels Etablierung einer Ressource zur funktionellen Genomanalyse des exprimierten (‘EXPRESSed‘) Anteils des Roggengenoms (GABI-RYE EXPRESS)“ (Teilprojekt A)

b) From RNA-seq to large-scale genotyping – genomics resources for rye (*Secale cereale* L.)

4. Autor(en) [Name(n), Vorname(n)]	5. Abschlussdatum des Vorhabens
a) Bauer, Eva; Haseneyer Grit b) Haseneyer, Grit; Schmutzer, Thomas; Seidel, Michael; Zhou, Ruonan; Mascher, Martin; Schön, Chris-Carolin; Taudien, Stefan; Scholz, Uwe; Stein, Nils; Mayer, Klaus F. X.; Bauer, Eva	Juni 2011 6. Veröffentlichungsdatum September 2011 7. Form der Publikation Fachzeitschrift
8. Durchführende Institution(en) (Name, Adresse)^ Technische Universität München, Lehrstuhl für Pflanzenzüchtung, Emil-Ramann-Str. 4, 85354 Freising Bioinformatik und Informationstechnologie, Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK), Corrensstr. 3, 06466 Gatersleben Genomdiversität, IPK, Corrensstr. 3, 06466 Gatersleben MIPS/IBIS, Institut für Bioinformatik und Systembiologie, HelmholtzZentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg	9. Ber. Nr. Durchführende Institution keine 10. Förderkennzeichen 0315063A 11. Seitenzahl a) 22 b) 13
12. Fördernde Institution (Name, Adresse) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. Literaturangaben 2 14. Tabellen 3 15. Abbildungen 8

16. Zusätzliche Angaben
keine

17. Vorgelegt bei (Titel, Ort, Datum)
nein

18. Kurzfassung

Zu Beginn des Projekts standen für Roggen etwas mehr als 9.000 EST Sequenzen in öffentlichen Datenbanken, ca. 200 PCR-basierte Marker und eine Handvoll genetischer Karten mit geringer Markerabdeckung zur Verfügung. Die Entwicklung zuverlässiger und kosteneffizienter Hochdurchsatz-Sequenziertechnologien und -Genotypisierungsplattformen ermöglichte die Untersuchung und Analyse eines Pflanzentranskriptoms in absehbarer Zeit und mit vertretbarem monetärem Aufwand.

Hauptziel des Verbundprojekts GABI RYE-EXPRESS war es, grundlegende Ressourcen für die Genomanaylse in Roggen (*Secale cereale* L.) zu schaffen.

Fünf Roggeninzuchtlinien wurden für die experimentellen Arbeiten herangezogen. Verschiedene RNA Proben je Inzuchtlinie lieferten eine diverse und umfangreiche Ansammlung von Roggentranskripten für die Sequenzierung. Die Sequenzierung von fünf Linien ermöglichte die Detektion von Sequenzpolymorphismen, auf deren Grundlage sich ein Genotypisierungsarray erstellen ließ.

Nach der Assemblierung von 2.5 Mio. Sequenzreads aus der Roche 454 GS FLX Sequenzierung wurde eine umfangreiche Genomressource erstellt, die 115.400 EST Sequenzen, einen Genotypisierungsarray mit 5.234 SNPs sowie eine hochdichte Transkriptkarte mit 3.562 SNPs und 211 SSRs umfasst und der wissenschaftlichen Gemeinschaft frei zugänglich ist. Durch die Genotypisierung von Kartierungspopulationen mittels des SNP-Arrays konnte eine Transkriptkarte für Roggen erstellt werden.

Die Ergebnisse aus dem GABI RYE-EXPRESS Projekt stellen für die weitere Forschung im Roggen eine fundamentale Grundlage dar und sind für die vergleichende Genomik der Triticeae von hohem Interesse. Mit der Sequenzressource und dem RYE5k-SNP Array wurden zudem molekulare Tools geschaffen, die nachfolgende Forschungsprojekte in Roggen erheblich voranbringen.

19. Schlagwörter
Roggen, Transkriptom, Sequenzierung, Ressource, Genotypisierung, SNP-Array

20. Verlag	21. Preis
BioMed Central, Floor 6, 236 Gray's Inn Road, London, UK	1.365,00 Euro

I. KURZDARSTELLUNG

1. Aufgabenstellung

Hauptziel des Verbundprojekts GABI RYE-EXPRESS war es, grundlegende Ressourcen für die Genomanaylse in Roggen (*Secale cereale* L.) zu schaffen. Fünf Roggeninzuchtlinien wurden für die experimentellen Arbeiten herangezogen. Verschiedene RNA Proben je Inzuchtlinie sollten eine diverse und umfangreiche Ansammlung von Roggentranskripten für die Sequenzierung liefern. Die Sequenzierung von fünf Linien ermöglicht die Detektion von Sequenzpolymorphismen, auf deren Grundlage sich ein Genotypisierungsarray erstellen lässt. Dieser Array kann für die Genotypisierung beliebig vieler Individuen verwendet werden. Durch die Genotypisierung von Kartierungspopulationen oder diverser Roggenpopulationen sollten eine Transkriptkarte für Roggen erstellt bzw. populationsgenetische Parameter und das genomweite Kopplungsphasenungleichgewicht im Roggen geschätzt werden. Durch die Erstellung umfangreicher genomischer Ressourcen in GABI RYE-EXPRESS kann Deutschlands Position in der Roggenforschung und -züchtung im weltweiten Wettbewerb gestärkt werden.

2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

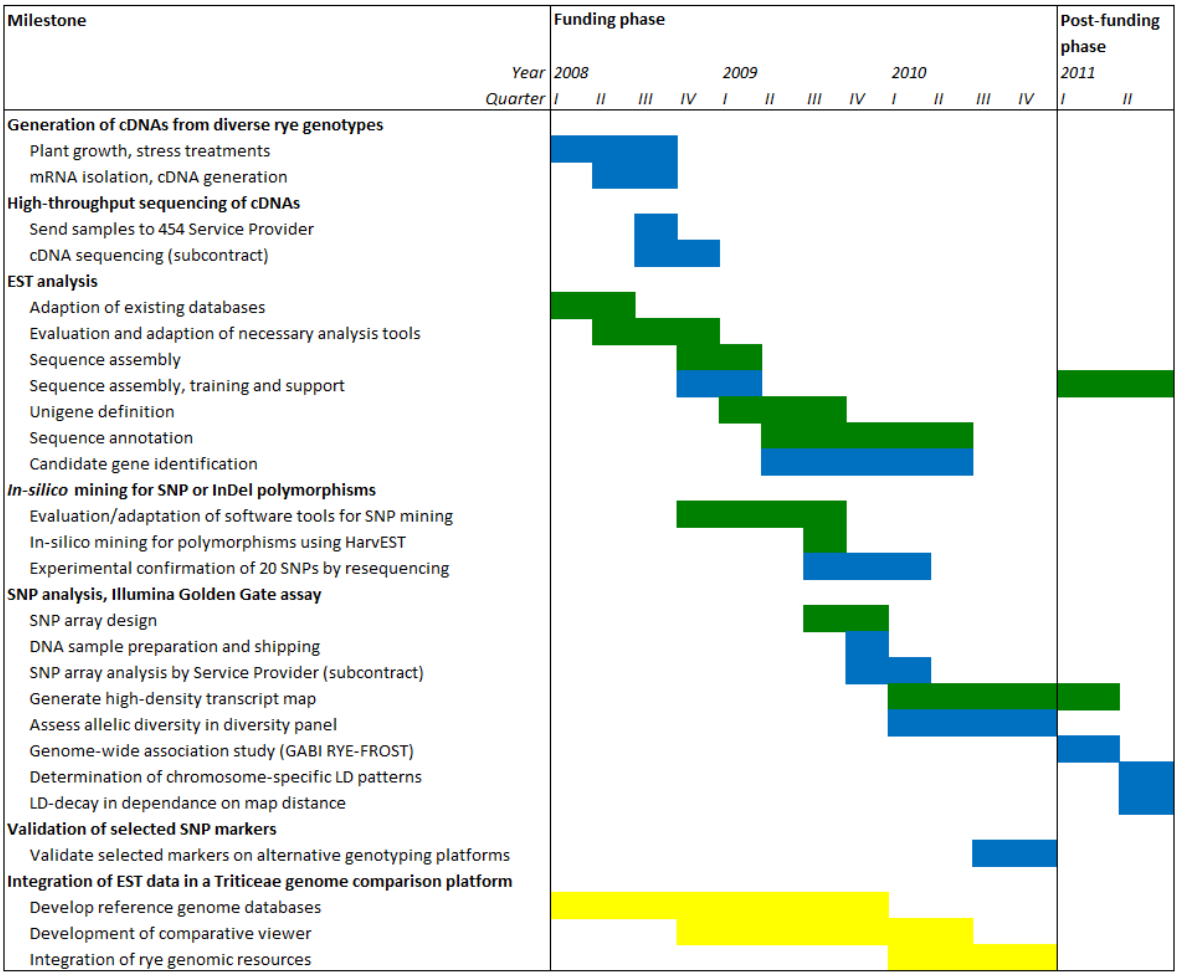
Der Bewilligungszeitraum von GABI RYE-EXPRESS fiel in eine Phase der schnelllebigen Entwicklung von Sequenzierungstechnologien und Genotypisierungsverfahren. Aufgrund des raschen Fortschritts in diesen Bereichen boten sich neue Möglichkeiten für die Erstellung einer umfangreichen Sequenz- und Genotypisierungsressource für Roggen. Dank der vorhandenen Expertise bei den Projektpartnern bezüglich Next-Generation-Sequencing (NGS), Genomanalyse, vergleichender Genomik und Bioinformatik wurden große Fortschritte für den Bereich Ressourcenentwicklung und Genomforschung im Roggen erwartet.

3. Planung und Ablauf des Vorhabens

Innerhalb des Projektes sind Planung und Ablauf durch einen Balkenplan (Tab 1) beschrieben, aus dem die Aufgaben der Projektpartner IPK, HMGU und TUM während des Bewilligungszeitraums hervorgehen.

Tabelle 1 Planung des Projektablaufs

Arbeiten während des Bewilligungszeitraums sind in grün (IPK), gelb (HMGU) und blau (TUM), Arbeiten während der Projektverlängerung sind in hellgrün (IPK) und hellblau (TUM) dargestellt.



4. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Zu Beginn des Projekts standen für Roggen etwas mehr als 9.000 EST Sequenzen in öffentlichen Datenbanken, ca. 200 PCR-basierte Marker und eine handvoll genetischer Karten mit geringer Markerabdeckung zur Verfügung. Die Entwicklung zuverlässiger und kosteneffizienter Hochdurchsatz-Sequenziertechnologien und -Genotypisierungsplattformen ermöglichte die Untersuchung und Analyse eines Pflanzentranskriptoms in absehbarer Zeit und mit vertretbarem monetärem Aufwand. Die aus diesem Projekt erwarteten wissenschaftlichen Ergebnisse bilden die Grundlage für die Etablierung des Roggens als Modellpflanze u. a. zur Untersuchung der Toleranz gegenüber abiotischem Stress in Gräsern.

5. Zusammenarbeit mit anderen Stellen

Die Zusammenarbeit der Projektpartner innerhalb des GABI RYE-EXPRESS Projekts war äußerst konstruktiv. GABI RYE-EXPRESS vereint die wissenschaftliche Expertise auf den Gebieten der Molekularbiologie, Bioinformatik und Genomanalyse. TUM war im Rahmen des Verbundes verantwortlich für die Erstellung und Bereitstellung des Sequenziermaterials, die Abwicklung der Unteraufträge für Sequenzierung und Genotypisierung und die Zusammenführung der Ergebnisse aller Projektpartner. Außerdem hat TUM die molekularen Analysen durchgeführt und die Daten zur genomweiten SNP-Analyse geliefert. Eine interaktive Zusammenarbeit zwischen TUM und IPK bestand bei der Entwicklung und Zusammenstellung des SNP-Genotypisierungsarrays sowie bei der Erstellung der Transkriptkarte. IPK hat die Assemblierungen übernommen. In enger Kooperation zwischen TUM und HMGU wurden die vergleichenden Genomanalysen durchgeführt. Sämtliche Ergebnisse wurden unmittelbar nach ihrer Fertigstellung allen Projektpartnern zur Verfügung gestellt. Kurzzeitige Forschungsaufenthalte ermöglichten den direkten Austausch. Dringende Besprechungen konnten jederzeit in Telefonkonferenzen stattfinden. Projekttreffen fanden zweimal jährlich statt, um Ergebnisse auszutauschen und den weiteren Projektverlauf abzusprechen.

Während des gesamten Bewilligungszeitraums war die sehr enge Zusammenarbeit nicht nur projektintern zwischen den Teilvorhaben sondern auch projektübergreifend zwischen den Projekten des Genomics-Netzwerk (Abb. 1) gewährleistet. Die Koordination des Verbundprojekts durch TUM gewährleistete die enge Vernetzung mit anderen Roggenprojekten im Rahmen von GABI-FUTURE (GABI RYE-FROST, GABI-TILL / Teilprojekt Roggen). Im Bewilligungszeitraum konnte die gewonnene Sequenzinformation für den Fortgang der Projekte GABI-TILL und GABI RYE-FROST genutzt werden. Basierend auf den Sequenzen der 454 Sequenzierung wurden für Kandidatengene Oligonukleotide entwickelt, die eine vollkommene Sequenzähnlichkeit zu den Genotypen der TILLING sowie der mittel- und osteuropäischen Populationen gewährleisten. Zuvor wurden Oligonukleotide basierend auf homologen Genen anderer Gräser entwickelt, deren Sequenzähnlichkeit oft nicht ausreichte, um spezifische Amplifikate in der PCR zu erhalten. Für das Projekt GABI RYE-FROST konnten zusätzlich wertvolle Informationen zu genomweitem LD zu Verfügung gestellt wurden.

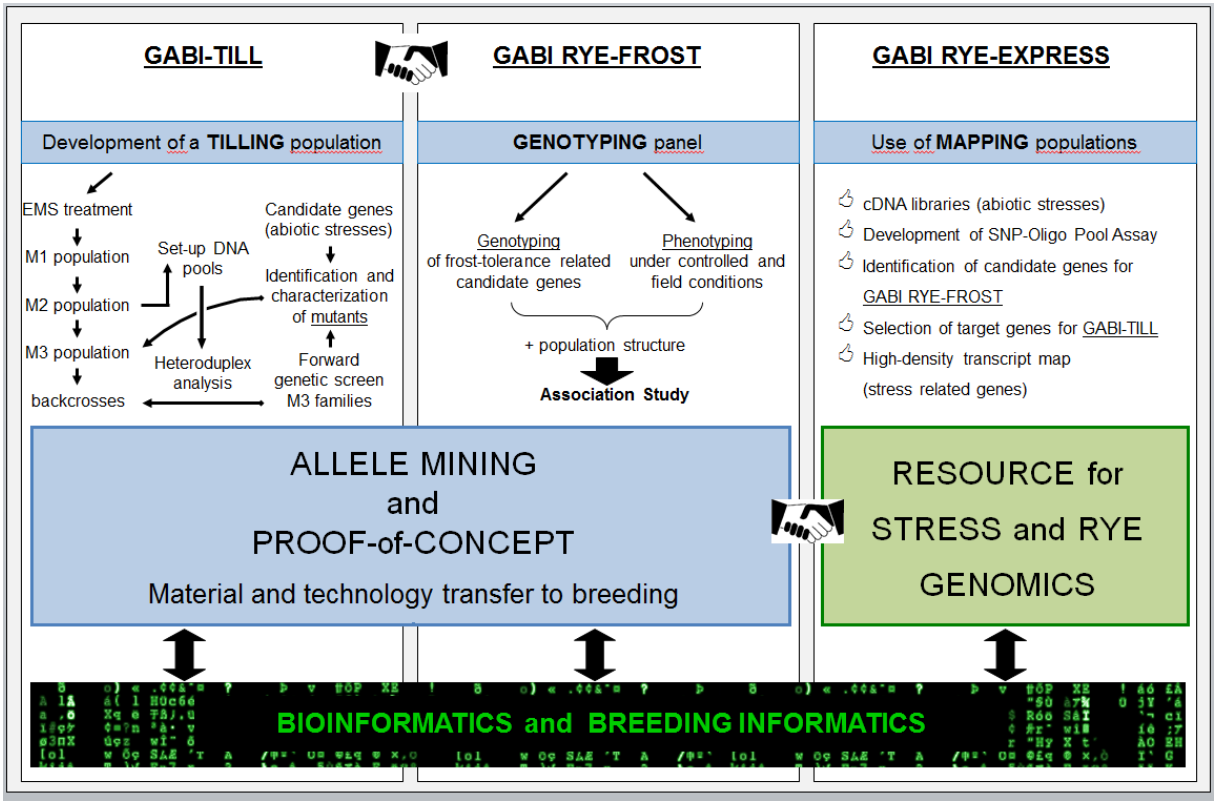


Abbildung 1 Vernetzung der Roggenprojekte im Rahmen von GABI FUTURE

Die Verbundpartner des GABI RYE-EXPRESS Projekts sind in nationale und internationale Forschungsaktivitäten eingebunden, wie z.B. weitere GABI-Projekte, die European Triticeae Genome Initiative (ETGI) und die COST Activity TRITIGEN. Dies sicherte die Anbindung an laufende Aktivitäten in der Genomanalyse bei Getreiden und bot die Möglichkeit, auf neue wissenschaftliche Entwicklungen reagieren zu können.

II. EINGEHENDE DARSTELLUNG

1. Verwendung der Zuwendung und der erzielten Ergebnisse

Für das Projekt GABI RYE-EXPRESS wurden fünf, genetisch diverse Roggeninzuchtlinien (Lo152, Lo225, Lo7, P87, P105) ausgewählt. Die Genotypen stammen aus den Regionen Deutschland und Russland und gewährleisten eine hohe genetische Diversität. Vier Linien sind Eltern von Kartierungspopulationen (Lo225 × Lo7 sowie P87 × P105). Durch die Sequenzierung und das Detektieren von Polymorphismen zwischen den Kreuzungseltern wird die Grundlage zur Erstellung von Transkriptkarten mit hoher Markerdichte geschaffen. Lo152 wurde als Kreuzungselter im Verbundprojekt GABI RYE-FROST eingesetzt. Die in GABI RYE-EXPRESS erlangten Sequenzen für diesen Genotyp dienen zum Auffinden neuer, mit Frosttoleranz im Zusammenhang stehender Kandidatengene.
Die Anzucht von je 20 Pflanzen erfolgte im Gewächshaus. Gewebeproben unterschiedlicher Entwicklungsstadien und verschiedener Stressbehandlungen wurden geerntet und Gesamt-RNA extrahiert. Die isolierten Gesamt-RNA Proben wurden zu gleichen Mengen für die cDNA Synthese bereitgestellt. Die Sequenzierergebnisse der zwei vorangestellten Testproben (Lo152, Lo225) ergaben, dass die cDNA Fraktionierung optimiert werden musste. Mit einer, von der vertis Biotechnologie AG neu fraktionierten cDNA Probe (Lo152), diesmal im Fragmentgrößenbereich von 500-700 bp, wurde ein weiterer 454-Testsequenzierlauf gestartet. Es war gelungen, mehr als 50% der Reads mit einer Länge von >200bp zu sequenzieren und kurze Fragmente (<100 bp) deutlich zu reduzieren (Tab 1). Die Anzahl der Reads pro Testsequenzierlauf war wie erwartet.

Tabelle 1. Ergebnisse (Anzahl Readsequenzen, Median, % Readsequenzen <100bp) der 454-Sequenzierungen der Testprobe Lo152 vor und nach optimierter cDNA-Fraktionierung

Probe	Anzahl Readsequenzen	Median [bp]	Readsequenzen <100 bp [%]*
Lo152 (vor)	31.808	157	30,05
Lo152 (nach)	32.839	206	6,73

* Readsequenzlänge nach clippen von 454 Adapter- und cDNA Syntheseprimer-Sequenz

Nach erfolgter Optimierung der Methode wurden die finalen cDNA Proben synthetisiert, in einem Größenbereich von 600-800 bp selektiert und zur Hochdurchsatz-Sequenzierung an das Fritz Lipmann Institut (FLI, Jena) versandt. Für jede der finalen Proben wurde erneut ein Testsequenzierlauf (je 1/16 Sequenzierbahn) durchgeführt, um die Qualität der Proben, bestimmt durch ‚erreichte Leseweite' und ‚BLAST Analyse', zu prüfen. Die BLAST Analyse

gegen verfügbare Datenbanksequenzen von Reis, Sorghum, Weizen, Gerste und *Brachypodium* zeigte für 79% der Readsequenzen einen Hit, 16% konnten bei der Analyse nicht bestimmt werden, 3% ergaben Treffer zu genomischen Sequenzen von Gerste, die verbleibenden 2% sind repetitive Sequenzen. Im Anschluss an die Qualitätsprüfung wurde die Hochdurchsatz-Sequenzierung der fünf Roggen-cDNAs durchgeführt (Abb. 2).

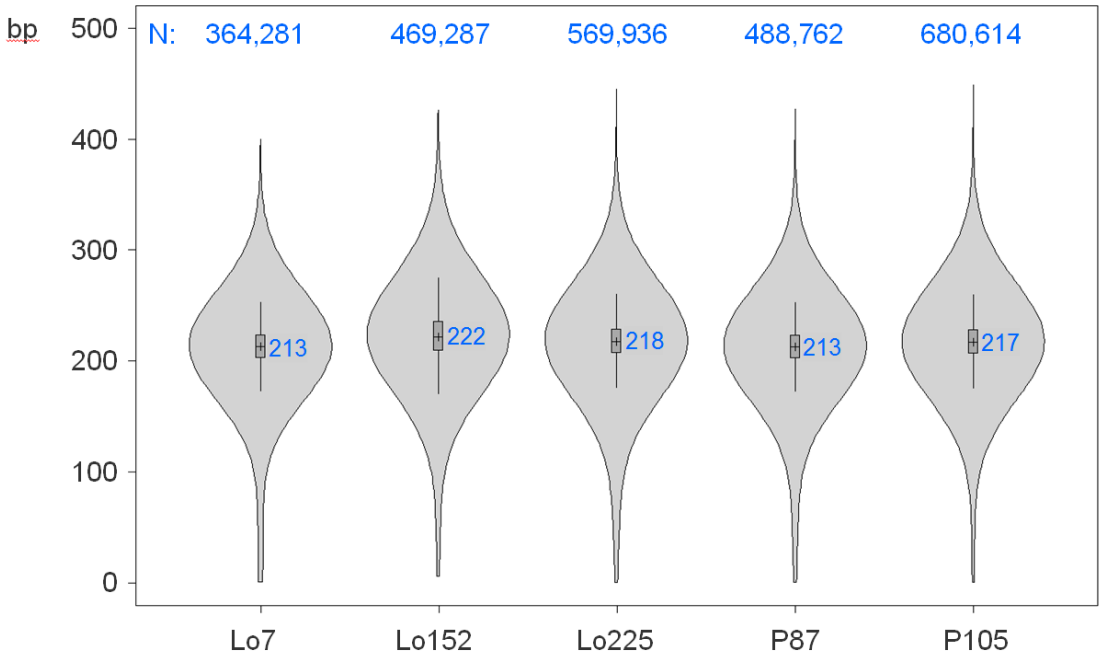


Abbildung 2. Violinenplots der finalen cDNA Proben nach erfolgtem 454-Sequenzierlauf je Inzuchtlinie; +: Median, N: Anzahl Sequenzierreads je Inzuchtlinie

Insgesamt wurden bei den fünf Sequenzierläufen etwa 2,5 Mio. Reads mit einer mittleren Readlänge von 216,6 bp generiert. Öffentlich verfügbare Algorithmen zur Auswertung und Assemblierung von 454 Sequenzdaten wurden herangezogen, um einerseits eine umfangreiche EST-Ressource zu erstellen und andererseits eine zuverlässige SNP Detektion innerhalb der assemblierten Sequenzen zu ermöglichen. Durch die interaktive Zusammenarbeit mit den Projektpartnern IPK und HMGU ist eine Strategie für die Datenauswertung erstellt worden (Abb. 3). Die zweckgebundene Assemblierung der 454 Readsequenzen, nach dem Entfernen von Sequenzieradaptern, cDNA-Syntheseprimern und Sequenzbereichen von minderer Qualität, ermöglichte eine rationale Vorgehensweise im Hinblick auf Rechenzeit, Rechnerauslastung und Speicherkapazität.

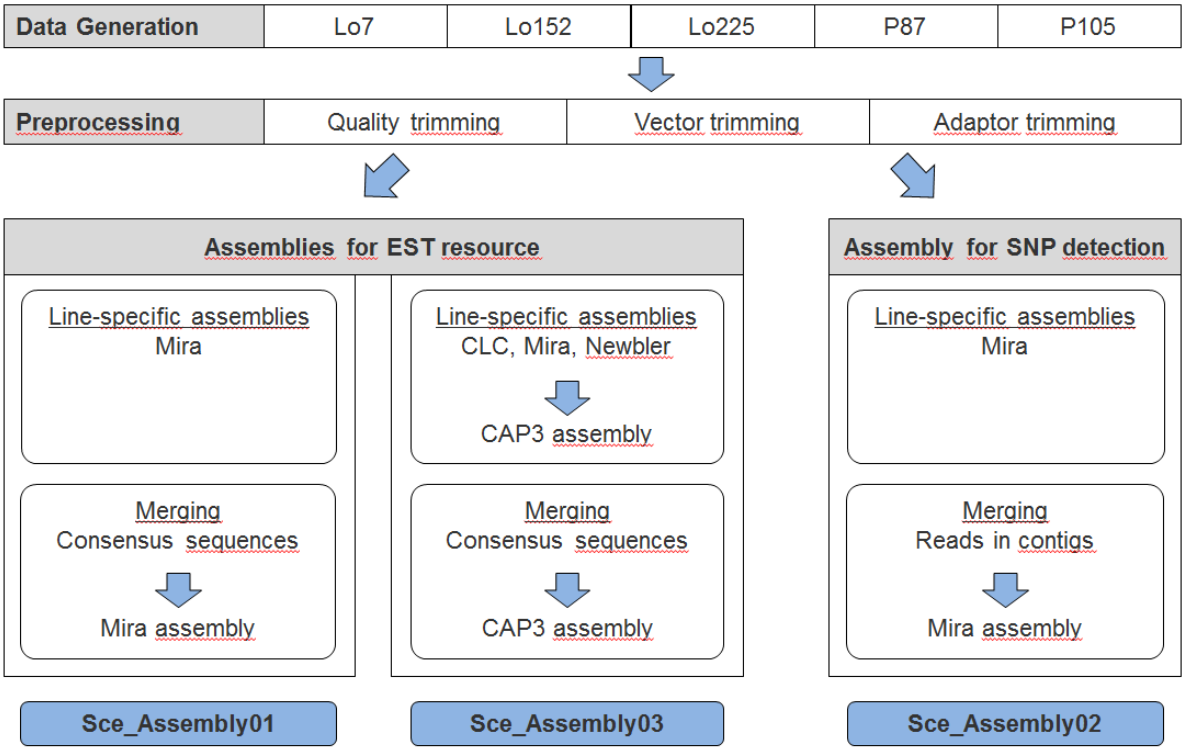


Abbildung 3 Illustration der zweckgebundenen Assemblierungsstrategien zur Erstellung einer EST Ressource und einer verlässlichen SNP-Detektion.

Das Sce_Assembly01 umfasst 87.199 multi-line Contigs, welche Readsequenzen aus zwei oder mehr Linien enthalten, 75.409 single-line Contigs, die Readsequenzen von nur einer Linie enthalten und 89.434 Singletons. Das Sce_Assembly03 beinhaltet 33.352 multi-line Contigs und 82.048 single-line Contigs während das für die SNP-Detektion erstellte Sce_Assembly02 138,339 Contigs umfasst. Die drei unterschiedlichen Assemblies bieten die Möglichkeit verschiedene down-stream Analysen zu adressieren. Das Sce_Assembly01 basiert auf einer Assemblierungssoftware und stellt die umfangreichste Sequenzressource dar. Allerdings können Artefakte bei der Assemblierung nicht ausgeschlossen werden. Diesen Nachteil konnten wir mit dem Sce_Assembly03 umgehen, bei dem nur Contigs berücksichtigt wurden, die von drei unabhängigen Assemblern erstellt wurden. Das Sce_Assembly02 dient der SNP-Detektion und erlaubt durch den Erhalt der Readinformation die Bestimmung der Coverage an einer ausgewählten SNP-Position. Die Erstellung dieses Assemblies war jedoch sehr zeit- und rechenintensiv, da im Gegensatz zu Sce_Assembly01 und SceAssembly03 keine Contigsequenzen beim Zusammenführen der Liniensequenzen verwendet wurden. Die zweckgebundene Assemblierung bietet somit den Vorteil, sinnvolle Kriterien aufzustellen und effizient mit der Datenmenge der NGS umzugehen.

Alle drei Assemblies sowie die linien-spezifischen Assemblies sind auf der GABI Primary Database Webseite herunterladbar (www.gabipd.org).

Die Charakterisierung des erstellten Assemblies erfolgte durch BLAST Analysen gegen öffentlich verfügbare genomische Sequenzen von *Brachypodium*, Mais, Sorghum und Reis sowie gegen EST und Volllängen-cDNA Sequenzen von Gerste und Weizen (Abb. 4). Diese Analysen zeigen eine hohe Ähnlichkeit zu Sequenzen von Gerste, Weizen und *Brachypodium* gefolgt von Sorghum, Reis und Mais. Dies spiegelt die Phylogenie unter den Gräsern wider.

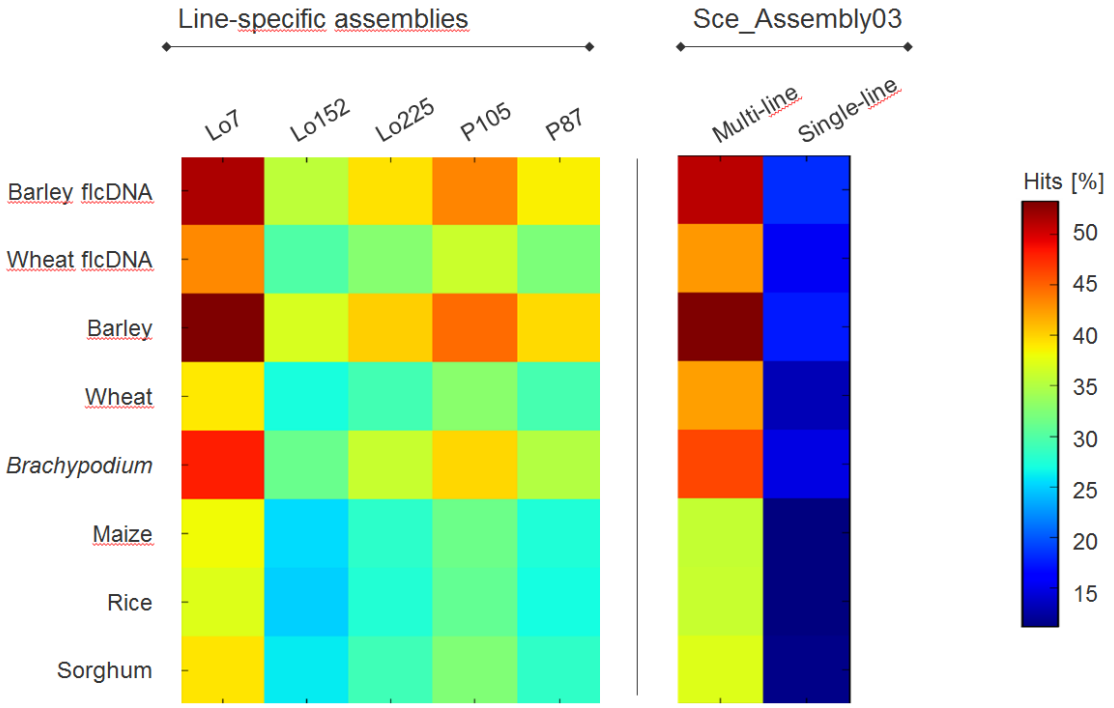


Abbildung 4 *Heatmap der BLAST Analyse der linienspezifischen Assemblies und des Sce_Assembly03 gegen öffentlich verfügbare Volllängen-cDNA Sequenzen und EST-Sequenzen von Gerste und Weizen sowie genomische Sequenzen von Brachypodium, Mais, Sorghum und Reis. Die Skala zeigt % Hits mit den entsprechenden Sequenzen.*

Ein wesentlicher Punkt bei der SNP Detektion war das Festlegen von Qualitätskriterien. Im Folgenden sind Qualitätsparameter aufgelistet, die eine sichere SNP Detektion gewährleisteten und das Filtern von SNPs in Qualitätskategorien ermöglichten:

- Genotypspezifische Analyse der Nukleotidfrequenzen an der SNP Position (χ^2-Test): inter- versus intragenotypische Polymorphismen
- Position innerhalb des Contigs, Berücksichtigung der Exonstrukturen
- Information zur Allelie der SNPs (bi-, tri-, tetraallelisch)
- Notwendigkeit der Implementierung mehrerer *in-silico* SNP Detektions-Tools zur Absicherung der identifizierten Polymorphismen (z.B. QualitySNP, AutoSNP)
- Nähe zu Homopolymeren

- BLAST Suche der assemblierten Contigs zur
  - ◊ Identifikation von Contigs, die einem Gen entsprechen
  - ◊ Lokalisierung des Contigs innerhalb des Gens (konservierte Domänen)
  - ◊ Gewinnung von Informationen über Genfamilien, Auftreten von Paralogen und Genduplikationen in anderen Gräsern

Anhand dieser erhobenen Qualtitätsparameter je SNP war ein schrittweises Filtern der SNPs nach bestimmten Parametern bzw. Parameterkombinationen möglich.
Zur Verifikation der im Sce_Assembly02 erstellten Contigs und der darin vorhergesagten SNPs wurde eine Re-Sequenzierung mittels Sanger-Methode durchgeführt. Hierfür wurden für 20 Contigs und 5 Genotypen 175 Sequenzen (13 single reverse, je 3 non-assembled forward und reverse, 156 assembled forward und reverse) ausgewertet. Insgesamt konnten 295 SNPs (180 Exon + 115 Intron) in ca. 16,13 kb (11,62 kb Exon + 4,51 kb Intron) detektiert werden, d.h. 1 SNP/54 bp (1 SNP/64 bp Exon; 1 SNP/39 bp Intron).
Bei der Auswertung des *de novo* Sce_Assemblies02 in den entsprechenden Contigs wurden 177 potentielle SNPs visuell ermittelt, von denen 143 (80,8%) mit den SNPs aus der Re-Sequenzierung übereinstimmen, demzufolge 34 nicht detektiert wurden. Der Grund hierfür: Die Contigs enthalten keine Readinformation von dem Genotyp, der das andere SNP Allel trägt. Somit zeigen die 34 Contigs in dem Sce_Assembly02 an den vermeintlichen SNP Positionen nur ein Allel. Für den Vergleich von *de novo* Assembly und Re-Sequenzierung und anschließender Validierung der SNP-Detection-Pipepline (=RYEpline) kann das Ergebnis wie folgt zusammengestellt werden (Tab 2):

Tabelle 2. Ergebnis der experimentellen SNP Validierung

Ergebnis		**Anzahl SNPs**
1)	Im Assembly vorhanden, bestätigt durch Re-Sequenzierung, detektiert in RYEpline	66
2)	NICHT im Assembly vorhanden, gefunden durch Re-Sequenzierung, NICHT detektiert in RYEpline	63
3)	Im Assembly vorhanden, bestätigt durch Re-Sequenzierung, NICHT detektiert in RYEpline	58
4)	Im Assembly vorhanden, NICHT bestätigt durch Re-Sequenzierung, detektiert in RYEpline	21
5)	Andere Ursachen	6

Es stellt sich die Frage, wie viele der im *de novo* Assembly vorhandenen SNPs mit der RYEpline detektiert werden konnten. Für die gewählten 20 Contigs finden sich 119 SNPs im Output der RYEpline, von denen 89 in den Bereich der re-sequenzierten Amplicons fallen.

Von diesen 89 sind 66 durch die Re-Sequenzierung bestätigt worden. Die verbleibenden 23 SNPs, die in der RYEpline detektiert wurden, aber nicht durch die Re-Sequenzierung bestätigt werden konnten, finden sich zwar im *Sce_Assembly02* wieder, kommen aber durch Misassemblierung (N=9), Sequenzierfehler in der Anfangssequenz (N=5) und intraline SNPs (N=5) zustande (Rest N=2 unklar). Vermutlich lassen sich solche Kandidaten aufgrund ihrer Position im Contig (Nähe zum Anfang/Ende der Sequenz) und des „no intraline SNP" Kriteriums eliminieren.

Die aus der SNP-Validierung gewonnenen Erkenntnisse zeigen, dass die SNP-Detektion mittels verfügbarer bzw. modifizierter automatischer Algorithmen noch weiter optimiert werden kann. Aufgrund dessen wurde die Anzahl von 227.034 in GigaBayes detektierten SNP-Kandidaten durch Filterkriterien auf 17.917 reduziert (Tab. 3) und mit Kandidatengenen aus dem GABI RYE-FROST Projekt ergänzt. Diese SNP-Kandidaten wurden zum Einschätzen ihrer Eignung zur Genotypisierung (Final_Score) an die Firma Illumina geschickt.

Tabelle 3. Filterkriterien zur Reduktion der Anzahl SNP-Kandidaten für die visuelle Validierung

Filterkriterium[a]	**Verbleibende Contigs (total=277.033)**
SNP: exclude -/T, -/A, -/G, -/C, -/N, T/N, A/N, C/N, G/N, -/T/N, -/A/N, -/C/N, -/G/N, T, A, G, C	186.370
IUPAC in oligoframe: 0 and 1 accepted	50.738
Allele_type: bi-allelic accepted	48.092
QualityRatio: >90 accepted	43.392
Distance_to_homopolymere: >5bp accepted	23.253
Distance_to_INDEL: >60bp accepted	23.253
Population_polymorphy: exclude control and empty	21.702
Distance_to_ContigEnd: >60bp accepted	19.958
Repeatelement und/oder SpliceSite in Oligoframe	17.917

a: SNP – Ausschluss von InDels und nicht-eindeutigen SNP-Allelen, IUPAC in oligoframe – kein oder maximal ein SNP im Oligoframe, Allele-type - nur bi-allelische SNPs wurden berücksichtigt, QualityRatio – Prozentuales Verhältnis von Qualitätswert an der SNP Position zu den Qualitätswerten der umliegenden Basen >90, Distance_to_homopolymer - Abstand zum Homopolymer mindestens 5bp, Distance_to_INDEL - Abstand zum INDEL mindestens 60bp, Population_polymorphy - Polymorphismus tritt zwischen den Kartierungseltern auf, Distance_to_ContigEnd – Abstand zum Contigende mindestens 60bp, Repeatelement und/oder SpliceSite in Oligoframe – Ausschluss von SNPs sobald ein Repeatelement oder eine putative Splicingstelle im Oligoframe detektiert wurden

Anhand der Illumina Assay Design Scores ergab sich eine Reduzierung der SNP-Kandidaten durch Final_Score zwischen 0,95 – 1,00, dem Qualitätskriterium zum Probe-Design. Durch

diesen Filterschritt blieben 9.278 SNP-Kandidaten, die visuell evaluiert wurden. Aus diesen konnten 3.171 SNPs selektiert werden. Um die Anzahl SNPs zu erhöhen, wurde das Final_Score Kriterium auf 0,80-1,00 gelockert. Zusätzlich sind noch solche SNP-Kandidaten hinzugenommen worden, die aus einem Contig stammen. Diese geben bei der Kartierung nicht unbedingt eine Zusatzinformation, können aber zum Abschätzen von intragenischem LD genutzt werden.

Final wurden 5.234 SNPs für die Array-Produktion an Illumina gesendet und der RYE5k-SNP Array wurde ausgeliefert.

Aufgrund der methodischen Weiterentwicklung und der günstigeren Preise in der Genotypisierungstechnologie wurde in Absprache mit den Projektpartnern IPK und HMGU sowie dem Projektträger entschieden, den iSelect bead array der Firma Illumina für die Genotypisierung zu verwenden. Mit dieser Methode konnten im Vergleich zum GoldenGate Oligo Pool Assay bei kostenneutralem Aufwand bis zu 6,000 SNPs (+300%), statt vorher 1,536 SNPs sowie 1,152 (+50%), statt 768 Proben, genotypisiert werden (Abb. 5).

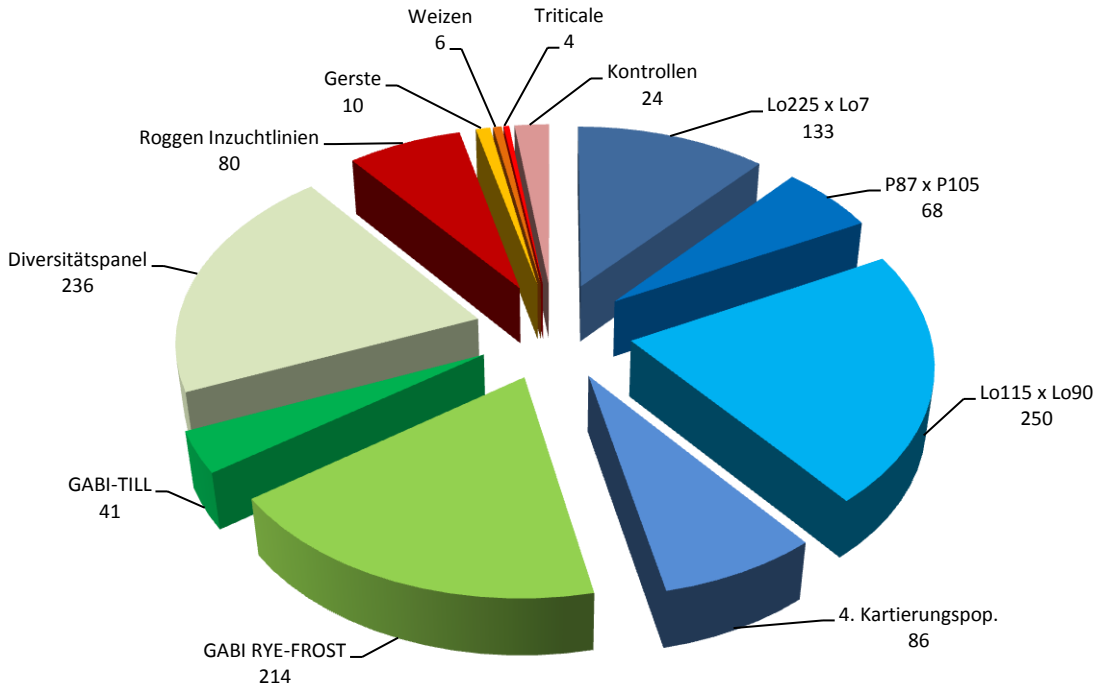


Abbildung 5 Zusammenstellung der Genotypisierungsproben für den iSelect bead array (blau: Kartierungspopulationen, Zahlen geben die Anzahl Genotypen wieder)

Insgesamt wurden Genotypisierungsdaten von vier Kartierungspopulationen zur Erstellung der hochauflösenden Transkriptkarte erhoben. Neben diesen wurden die Kollektionen aus

GABI RYE-FROST, das Linien-Sortiment aus GABI-TILL, diverse Roggeninzuchtlinien, ein Diversitätspanel und Kartierungseltern von Gerste (Oregon Wolfe Barley Population, Morex, Barke, Steptoe), Weizen (Dream, Chinese Spring, Mulgara) und Triticale (Modus, Saka3006) für die Genotypisierung ausgewählt. Die 1.152 DNA-Proben wurden photometrisch quantifiziert und auf Agarosegelen qualitativ überprüft. Nach Analyse der Genotypisierungsdaten und Erstellung von Clustern je SNP Assay für die unterschiedlichen Genotypklassen konnten insgesamt 4.557 der 5.234 SNP-Assays (87%) ausgewertet werden. Die Auswertung erfolgt mit dem GenTrain 2 Algorithmus, der in der Software GenomeStudio V2009.1, in der Komponente Genotyping v1.1.9 implementiert ist. Je SNP wird ein Mittelwert über alle Beads dieses SNPs angegeben. Für die Auswertung der Assays sind im Wesentlichen zwei Punkte interessant: die Call Rate und der GenTrain Score. Die Call Rate, auch Call Frequenz, bezieht sich auf die erfolgreich ausgewerteten Assays, liegt zwischen 0 und 1 und wird berechnet als #Calls / (#NoCalls + #Calls). Die Call Rates der genotypisierten Proben liegen zwischen 0,6 und 0,9. Der GenTrain Score gibt die Verlässlichkeit der SNP Detektion an, basierend auf der Zuordenbarkeit einer DNA in die Genotypenklassen AA, AB, BB. Dieses statistische Maß ist abhängig von den Clusterformen, die gebildet werden, und der relativen Clusterdistanz. Laut Illumina ist der GenTrain Score ein Wert, der die visuelle Auswertung der Assays durch einen Experten nachahmt. Dies bietet einen sehr guten Anhaltspunkt für unerfahrene Betrachter, muss aber durch genauere Prüfung validiert werden. Ein Problem für die Auswertung einiger SNPs mittels automatischem Clustering ist, dass nicht für jeden SNP die drei Genotypklassen AA, AB, BB vorhanden sind. In diesem Fall werden die fehlenden Cluster durch neuronale Netze geschätzt und sind somit fehlerbehaftet.

Für Akzessionen der Getreidearten Gerste, Weizen und Triticale ergaben 63, 76, bzw. 84 % der in Roggen auswertbaren SNP-Assays ein Signal. Es ergibt sich somit die Möglichkeit, den entwickelten Genotypisierungsarray auch für Weizen oder Triticale zu nutzen, bei denen die Erstellung von SNP-Arrays auf der illumina Plattform zurzeit noch nicht erfolgt ist. Für die Roggenpopulation Altevogt14160 und 54 Roggeninzuchtlinien wurde der Heterozygotiegrad bestimmt. Durch das Poolen von Einzelpflanzen der Population Altevogt14160 konnte gezeigt werden, dass mit einem Pool von acht Einzelpflanzen das Maximum der Heterozygotie in den betrachteten 16 Einzelpflanzen nahezu erreicht wird (Abb. 6). Für populationsgenetische Untersuchungen ist dies eine wichtige Information zur Anzahl von Einzelpflanzen bei der Erstellung von repräsentativen Stichproben für eine bestimmte Population. In den untersuchten Inzuchtlinien konnten zwischen 4.0 % und 20.5 % Restheterozygotie nachgewiesen werden. Ein höheres Ausmaß an Restheterozygotie trat in dem Pollenelterpool im Vergleich zum Saatelterpool auf, was auf den höhere Anzahl an Inzuchtgenerationen im Saatelternpool im Vergleich zum Pollenelterpool zurückzuführen ist.

Bei den *in silico* berechneten Werten der Restheterozygotie kann eine Überschätzung nicht ausgeschlossen werden, da in dem Assembly nicht zwischen Polymorphismen aufgrund von Heterozygotie und Polymorphismen aufgrund von Paralogenassemblierung unterschieden werden kann.

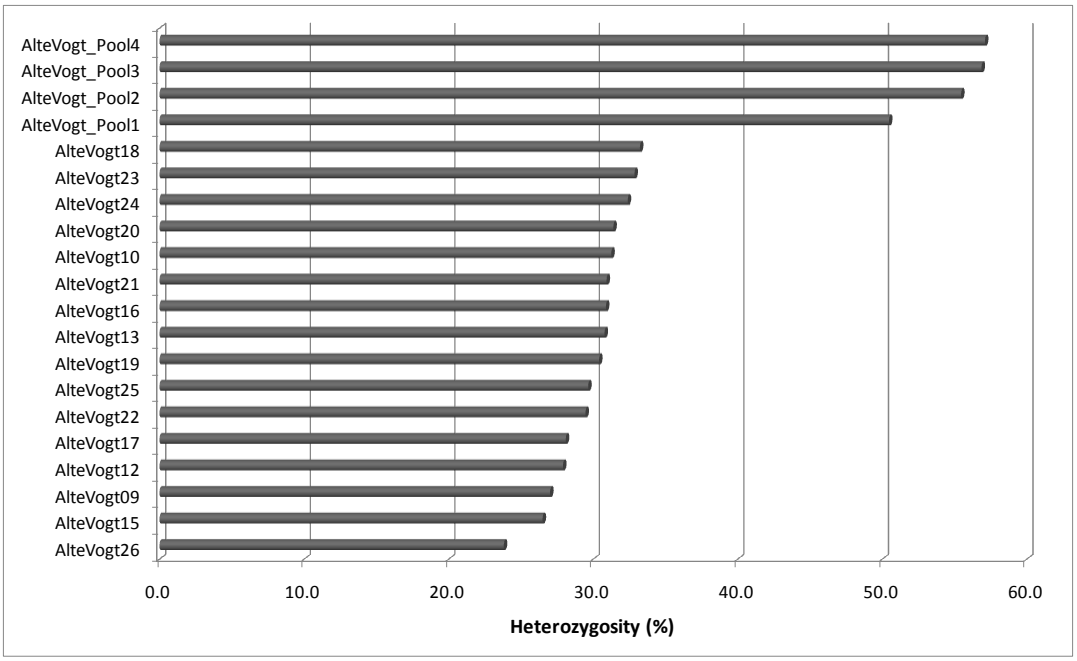


Abbildung 6 Heterozygotie in 16 Einzelpflanzen der Roggenpopulation Altevogt14160 und vier Pools aus je 2 (Pool1), 4 (Pool2), 8 (Pool3) und 16 (Pool4) Einzelpflanzen. Die kleineren Pools sind Teilmenge der jeweils folgenden Pools. Insgesamt wurden 2.124 in der Population polymorphe SNPs betrachtet.

Die Genotypisierungsdaten sind um ein vielfaches umfangreicher als bei Projektbeginn absehbar. Sie ermöglichen eine erste Schätzung des genomweiten LD in Roggen basierend auf SNP-Daten. Informationen der hochdichten Transkriptkarte, die vom Projektpartner IPK erstellt wurde, stellen die Grundlage für diese Arbeit dar. Basierend auf den Genotypisierungsdaten der vier Kartierungspopulationen und den detektierten SSR Markern wurde die Anzahl Polymorphismen je Chromosom ermittelt (Abb. 7). Zusätzlich wurden in der JKI Population (JKI1 × JKI2) Ankermarker kartiert, die eine Verbindung zu früher erstellten Roggenkarten ermöglichen.

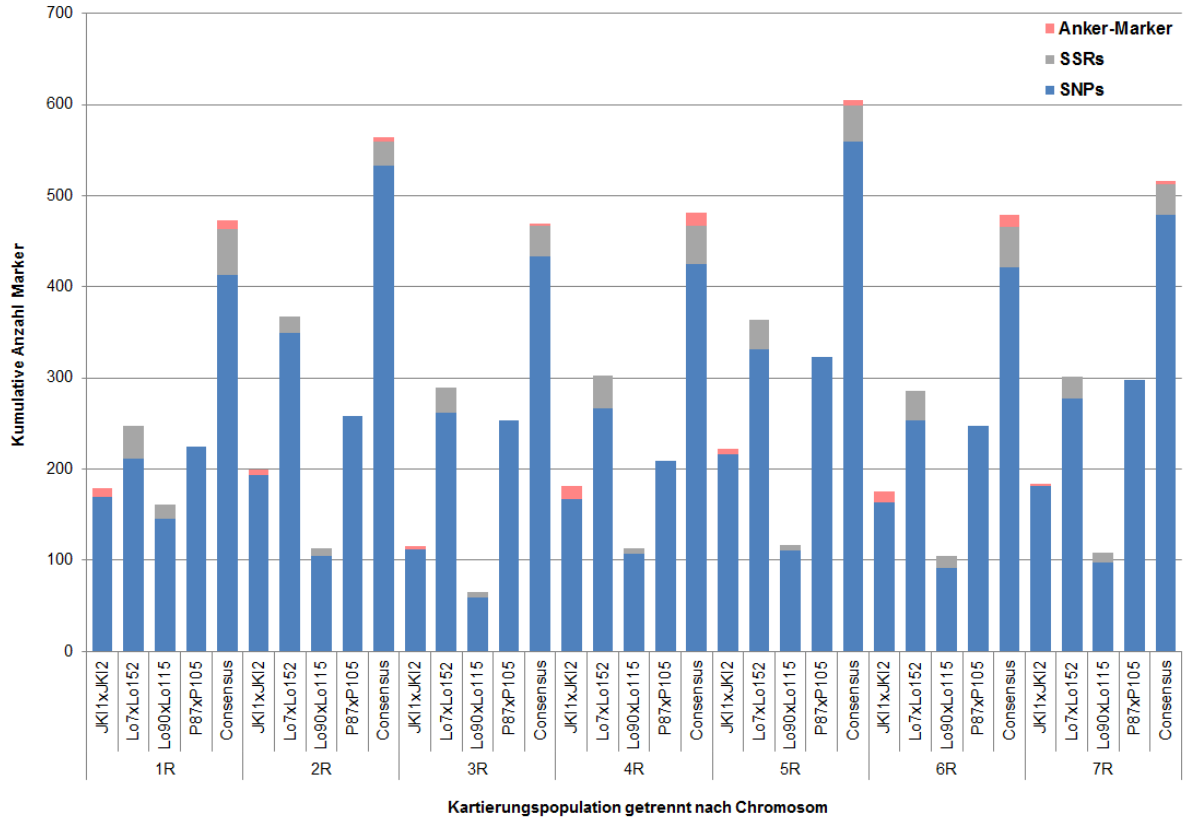


Abbildung 7 Kumulative Anzahl von SNP, SSR und Anker-Markern je Roggenchromosom in vier Kartierungspopulationen und in der Consensus-Karte dieser vier Populationen.

Insgesamt konnten 3,264 SNP, 271 SSR- und 53 Ankermarker in der Consensus-Karte auf die sieben Roggenchromosomen verteilt kartiert werden. Die hohe Datenmenge erfordert die Verwendung neuer Softwarepakete und die Etablierung von Skripten für die effiziente Verrechnung und anschauliche Darstellung der Ergebnisse. LD Studien sowohl genomweit als auch chromosomenweise für das genotypisierte Roggenmaterial ergaben einen schnellen Abbau des LD innerhalb von weniger als 1 cM. Beispielhaft ist dies in Abbildung 8 für die Kollektion aus GABI RYE-FROST gezeigt. Es wurden keine gravierenden Unterschiede zwischen Inzuchtlinien und Populationen beobachtet, was einerseits auf die noch unzureichende Markerdichte je Chromosom zurückzuführen ist. Zudem ist anzunehmen, dass sich in einzelnen Genomregionen aufgrund von Selektion ein abweichendes Bild ergibt.

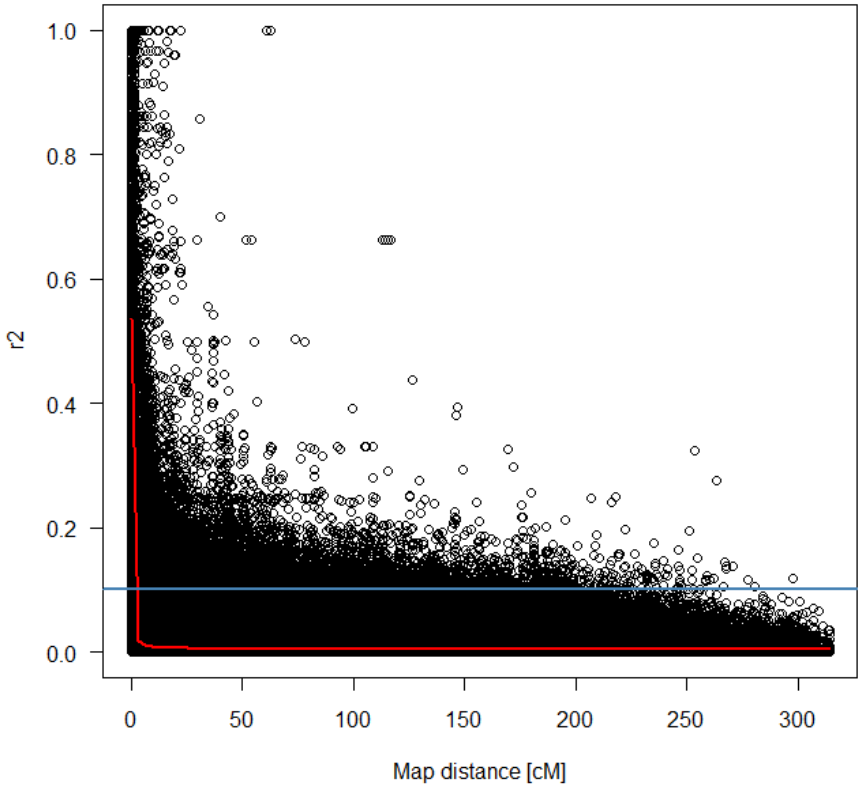


Abbildung 8 *Genomweites LD (r^2) in Abhängigkeit von der Kartendistanz (cM) in der GABI RYE-FROST Kollektion (N=201). Die rote Kurve zeigt den LD Abbau, die horizontale blaue Linie markiert den Wert r^2=0,1.*

2. Wichtige Positionen des zahlenmäßigen Nachweises

Siehe Anlage „Verwendungsnachweis Mittel“

3. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die Ergebnisse aus dem GABI RYE-EXPRESS Projekt stellen für die weitere Forschung im Roggen eine fundamentale Grundlage dar und sind für die vergleichende Genomik der Triticeae von hohem Interesse. Hinsichtlich des Wissensgewinns zu genomweitem LD, genetischer Diversität und Populationsstruktur im Roggen sind durch das Projekt GABI RYE-EXPRESS deutliche Fortschritte gemacht worden, die grundlegende wissenschaftliche Erkenntnisse auch für die zukünftige Roggenforschung liefern. Mit der Sequenzressource und dem RYE5k-SNP Array wurden zudem molekulare Tools geschaffen, die nachfolgende Forschungsprojekte in Roggen erheblich voranbringen. Demnach waren die geleisteten Arbeiten notwendig und angemessen und haben zu verwertbaren Ergebnissen geführt.

4. Voraussichtlicher Nutzen und Verwertbarkeit der Ergebnisse

Bereits im Verlauf des Berichtzeitraumes trat ein WPG Mitglied an das Konsortium heran mit dem Interesse vorzeitigen Zugang zu den Projektergebnissen zu erhalten. Zur Regelung dieser Angelegenheit wurde den WPG-Mitgliedern für 3 Monate exklusives Leserecht der Ergebnisse eingeräumt. Die Daten wurden zum Zeitpunkt der Manuskripteinreichung „Haseneyer G, Schmutzer T, Seidel M, Zhou R, Mascher M, Schön CC, Taudien S, Scholz U, Stein N, Mayer KF, Bauer E (2011) From RNA-seq to large-scale genotyping - genomics resources for rye (*Secale cereale* L.). BMC Plant Biology 11:13“ über GABI-PD den WPG-Mitgliedern zugänglich gemacht. Eine Nutzung des SNP-Arrays bereits vor der exklusiven Bereitstellung für die WPG-Mitglieder über GABI-PD wurde über ein Material Transfer Agreement (MTA) geregelt. Die Sequenzen und der SNP-Array können von WPG-Mitgliedern für eigene Forschungszwecke genutzt werden. Interessenten können die SNP-Liste zur Bestellung von Arrays bei Illumina für eigene Analysen nutzen.

Im Berichtszeitraum wurde ein MTA zwischen allen Projektpartnern und der Universität Aarhus unterzeichnet. Die mit diesem MTA bereitgestellten Sequenzen sind das Ergebnis einer BLAST-Analyse gegen die GABI RYE-EXPRESS EST Ressource unter Verwendung homologer Gensequenzen von NAC Transkriptionsfaktoren, Phytasen und BX Genen.

Die in GABI RYE-EXPRESS aus den Sequenzdaten entstandenen Assemblierungen Sce_Assembly01, Sce_Assembly02 und Sce_Assembly03 stehen seit Dezember 2010 den WPG Mitgliedern und seit April 2011 der Öffentlichkeit über GabiPD (www.gabipd.org) zur Verfügung.

Ein Verwertungsplan für die in GABI RYE-EXPRESS erhobenen Daten und erstellten Ressourcen wurde in Absprache mit den Projektpartnern angefertigt. Der Verwertungsplan gewährleistet, dass die etablierten Ressourcen für Roggen der öffentlichen Forschung auch durch Dritte und über die Projektlaufzeit hinaus zur freien Verfügung stehen.

5. Bekannt gewordene Fortschritte auf dem Gebiet des Vorhabens bei anderen Stellen

In der Softwareentwicklung ist bei dem MIRA Assembler eine neue Version veröffentlicht worden, woraufhin die Assemblierung der Readsequenzen zur Qualitätsverbesserung der Roggen EST-Ressource noch einmal durchgeführt wurde.

Die Weiterentwicklung von Assemblierungsstrategien, vorgestellt etwa in Kumar S, Blaxter ML (2010) Comparing *de novo* assemblers for 454 transcriptome data. BMC Genomics 11:571 wurden in die laufenden Analysen von GABI RYE-EXPRESS berücksichtigt.

6. Erfolgte und geplante Veröffentlichungen der Ergebnisse

Haseneyer G (2008) GABI RYE-EXPRESS-Unlocking the genetic potential of rye by establishing a functional genomics resource for the EXPRESSed portion of the rye genome. Talk, 8th GABI Status-Seminar March 04-06 2008, Potsdam, Germany

Bauer E, Haseneyer G, Mayer KFX, Pietsch C, Schön CC, Scholz U, Stein N (2008) GABI RYE-EXPRESS: Unlocking the genetic potential of rye by establishing a functional genomics resource for the EXPRESSed portion of the rye genome. Poster, 7th Plant Genomics European Meeting, September 24-27 2008, Albena, Bulgaria

Haseneyer G (2009) GABI RYE-EXPRESS: Unlocking the genetic potential of rye by establishing a functional genomics resource for the EXPRESSed portion of the rye genome. Talk, 9th GABI Status-Seminar, March 03-05 2009, Potsdam, Germany

Haseneyer G, Schmutzer T, Seidel M, Mayer KFX, Schön CC, Scholz U, Stein N, Bauer E (2009) Establishing a genomics resource for the EXPRESSed portion of the rye genome to unlock its genetic potential. Poster, 8th Plant Genomics European Meeting October 07-10, Lisbon, Portugal

Bauer E (2009) RYE-*EXPRESS:* Unlocking the genetic potential of rye by establishing a functional genomics resource for the EXPRESSed portion of the rye genome. Talk, 19th International Triticeae Mapping Initiative (ITMI) workshop, August 31-September 04 2009, Clermont-Ferrand, France

Haseneyer G, Schmutzer T, Seidel M, Mayer KFX, Schön CC, Scholz U, Stein N, Bauer E (2010) GABI RYE-EXPRESS: Using next-generation technologies for improvement of genetic and genomic resources in rye (*Secale cereale*). Poster, 10th GABI Status Seminar

Haseneyer G, Schmutzer T, Seidel M, Mayer K, Schön CC, Scholz U, Stein N, Bauer E (2010) GABI RYE-EXPRESS: A functional genomics resource for the EXPRESSed portion of the rye genome. Poster, GPZ Tagung–Innovations in Breeding Methodology, March 15-17 2010, Freising, Germany

Bauer E, Haseneyer G, Schmutzer T, Seidel M, Zhou R, Schön CC, Mayer KFX, Scholz U, Stein N (2010) GABI-RYE genomics resources: sequences, SNPs and beyond. Invited talk, Eucarpia Cereals Meeting, April 06-08 2010, Cambridge, UK

Haseneyer G, Schmutzer T, Seidel M, Zhou R, Schön CC, Scholz U, Mayer KFX, Stein N, Bauer E (2010) New genomics resources for rye (*Secale cereale* L.). Poster, GPZ Tagung–Genomics-based breeding, October 26-28 2010, Gießen, Germany

Bauer E, Haseneyer G, Schmutzer T, Seidel M, Zhou R, Schön CC, Mayer KFX, Scholz U, Stein N (2011) Bringing rye genomics on track: Transcripts, SNPs, maps, and diversity. Talk, Plant & Animal Genomes XIX Conference, January 15-19 2011, San Diego, CA, USA

Bauer E (2011) Rye genomics coming of age - new tools, new challenges. Talk, UMR INRA-UBP 1095, Génétique, Diversité et Ecophysiologie des Céréales, May 19 2011, Clermont-Ferrand, France

Haseneyer G, Schmutzer T, Seidel M, Zhou RN, Mascher M, Schön CC, Taudien S, Scholz U, Stein N, Mayer KFX, Bauer E (2011) From RNA-*seq* to large-scale genotyping - genomics resources for rye (*Secale cereale* L.). BMC Plant Biol 11

Martis M, Klemme S, Moghaddam AMB, Blattner FR, Macas J, Schmutzer T, Scholz U, Gundlach H, Wicker T, Šimková H, Novák P, Neumann P, Kubaláková M, Bauer E, Haseneyer G, Fuchs J, Doležel J, Stein N, Mayer KFX, Houben A (in Vorbereitung) The selfish B chromosome of rye evolved as a mosaic of multiple A chromosome and organelle derived sequences.

Martis M, Zhou R, Haseneyer G, Seidel M, Schmutzer T, Schön CC, Hackauf B, Scholz U, Bauer E, Stein N, Mayer KFX (in Vorbereitung) Five reciprocal transclocations shaped the *Secale cereale* genome after separation from a common *Triticeae* progenitor.

Auinger HJ, Bauer E, Li Y, Haseneyer G, Schön CC (in Vorbereitung) First genome-wide association study in rye and the comparison to genomic selection.

Haseneyer G, Bauer E, Schön CC, Geiger HH (in Vorbereitung) First survey of genetic variation, linkage disequilibrium, and population structure in rye (*Secale cereale* L.) based on genome-wide single nucleotide polymorphisms.