

Charakterisierung und Sequenzanalyse von T-DNA Insertionsstrukturen
und paralogen Bereichen im Genom von *Arabidopsis thaliana* basierend
auf Optimierungen der Daten-Erfassung und -Auswertung für die
GABI-Kat-Kollektion

Kumulative Dissertation

Zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften
(Dr. rer. nat.) an der Fakultät für Biologie der Universität Bielefeld

von

Nils Kleinbölting

April 2015

Nils Kleinbölting
Wickenkamp 2c
33615 Bielefeld
nkleinbo@cebitec.uni-bielefeld.de

Betreuer: Prof. Dr. Bernd Weisshaar

Inhaltsverzeichnis

1	Zusammenfassung	1
2	Einführung	3
2.1	Modellorganismus <i>Arabidopsis thaliana</i>	4
2.1.1	Das Genom von <i>A. thaliana</i>	4
2.1.2	Genomevolution in <i>A. thaliana</i>	5
2.1.3	Reverse Genetik und die GABI-Kat-Kollektion	9
2.2	Transformation durch <i>Agrobacterium tumefaciens</i>	10
2.2.1	<i>A. tumefaciens</i> und das Ti-Plasmid	10
2.2.2	Der Weg der T-DNA in den Zellkern	11
2.2.3	Integration in das Genom	15
2.2.4	Ort der T-DNA Insertion	17
2.3	Reparaturmechanismen für Doppelstrangbrüche	18
2.3.1	Homologe Reparatur	18
2.3.2	Nicht-homologe Reparatur	20
2.3.3	Mikrohomologie-abhängige Reparatur	21
2.4	Die Insertionslinienpopulation GABI-Kat	23
2.4.1	Transformation in GABI-Kat	23
2.4.2	Insertionsstellen-Vorhersage	24
2.4.3	Der Bestätigungs-Prozess (<i>Confirmation</i>)	26
2.4.4	GABI-Kat LIMS und SimpleSearch	29
2.4.5	Andere Kollektionen	32
3	Zielsetzung	34
4	Ergebnisse	35
4.1	Optimierung der Insertionsstellen-Vorhersage und des Umgangs mit Kontaminationen	36
4.2	Berechnung von Gruppen paraloger Gene für die Erzeugung von Doppelmutanten in <i>A. thaliana</i> und deren Bereitstellung	41
4.3	Entwicklung eines Primerdesigns optimiert für paraloge Bereiche des Genoms von <i>A. thaliana</i>	47
4.4	Bioinformatische Analyse von <i>Confirmation</i> -Sequenzen in Bezug auf Merkmale des Integrationsmechanismus	52
4.4.1	Untersuchung von <i>Border</i> -Schnittstellen	53
4.4.2	Untersuchung von Mikrohomologien und <i>Fillern</i>	53
4.4.3	Analyse der <i>Filler</i>	55

4.4.4	Analyse von Deletionen und Duplikationen an der Insertions- stelle	55
4.4.5	Größere Deletionen und Inversionen	56
4.4.6	Vergleich mit Daten von SALK-Linien	56
5	Zusammenfassende Diskussion	58
	Literaturverzeichnis	65
	Abkürzungsverzeichnis	75
	Publikation 1: GABI-Kat SimpleSearch: New features of the <i>Arabidopsis thaliana</i> T-DNA mutant database	76
	Publikation 2: GABI-DUPLO: a collection of double mutants to overcome genetic redundancy in <i>Arabidopsis thaliana</i>	82
	Publikation 3: An easy-to-use primer design tool to address paralogous loci and T-DNA insertion sites in the genome of <i>Arabidopsis thaliana</i>	98
	Publikation 4: Evaluation of the structural features of thousands of T-DNA insertion sites indicates a double-strand break repair based insertion mechanism	109

Zusammenfassung

Weist das Genom eines Organismus einen Defekt in einem bestimmten Gen auf, der dessen Funktion komplett aufhebt, spricht man von einer sogenannten *Knockout*-Mutante. Diese sind ein unentbehrliches Werkzeug für die reverse Genetik, denn aus der Beobachtung der Mutante lassen sich Rückschlüsse auf die Funktion des defekten Gens ziehen. Die GABI-Kat-Kollektion wurde zu dem Zweck entwickelt, Zugriff auf knapp 72.000 potentiell interessante *Knockout*-Linien in *Arabidopsis thaliana* anzubieten. Zur Generierung der Linien wurde die durch *Agrobacterium tumefaciens* vermittelte Transformation verwendet, bei der ein Teil des Tumor-induzierenden (Ti)-Plasmids, die T-DNA, in das Genom des Wirts-Organismus an einer zufälligen Stelle stabil integriert wird und aufgrund ihrer Größe zum vollständigen *Knockout* eines dort kodierten Gens führen kann. Ein großer Vorteil gegenüber anderen Methoden zur Generierung von *Knockouts* ist, dass die Sequenz der integrierten T-DNA bekannt ist. So lassen sich Primer erstellen, mit denen *Flanking Sequence Tags* (FSTs) sequenziert werden können, die zur Insertion benachbarte Sequenzen beinhalten. Durch Analyse der Sequenziererergebnisse lassen sich Rückschlüsse auf den genauen Ort der Insertionen ziehen. Forscher können eine GABI-Kat Linie bestellen, die eine für sie interessante Insertion enthält. Daraufhin wird über eine klärende PCR und nachfolgende Sequenzierung die Anwesenheit der Insertion bestätigt und die Samen schließlich an den Besteller versendet.

Im Rahmen dieser Arbeit wurden verschiedene Methoden innerhalb des GABI-Kat Projektes etabliert, um die Anzahl der bestätigten Insertionen zu erhöhen. Diese Methoden umfassten eine exaktere Vorhersage der FST-basierten Insertionsposition, die Einführung von Kontaminationsgruppen, um die richtige Linie unter identischen Vorhersagen in verschiedenen Linien zu identifizieren und die Aufklärung von schwerwiegenden Fehlern in der Zuordnung von FSTs. Außerdem wurden zwei Methoden zur Steigerung der Zuverlässigkeit von Analysen in paralogen Bereichen des Genoms entwickelt. Dies umfasste zum Einen paraloge Gruppen zur Aufklärung von Mehrdeutigkeiten in den Insertions-Vorhersagen und zum Anderen die Entwicklung eines Werkzeugs zur Generierung von eindeutigen Primern im Genom von *A. thaliana*. Diese Maßnahmen führten zu einem signifikanten Anstieg der Bestätigungsrate, die angibt, wie viele Insertionen experimentell bestätigt werden konnten von 78 % auf fast 88 %. Sie trugen maßgeblich dazu bei, dass möglichst viele Forscher die für sie interessanten Linien erhalten und unterstützten damit den wissenschaftlichen Informationsgewinn.

Aufgrund von Duplikationen und daraus resultierenden paralogen Genen im Ge-

nom von *A. thaliana* führen deren *Knockouts* in vielen Fällen nicht zwangsläufig zu einem sichtbaren Phänotyp. Durch mindestens eine zweite intakte Kopie, die die Funktion des defekten Gens teilweise oder komplett übernehmen kann, bleiben mögliche Auswirkungen des *Knockouts* oft unbemerkt. Um dieses Problem zu adressieren, wurde in dieser Arbeit eine Liste homologer Gene in *A. thaliana* berechnet anhand derer überprüft werden kann, ob der *Knockout* mehrerer Gene bei der Untersuchung eines Gens erforderlich ist. Anhand dieser Liste wurden 200 Doppelmutanten für vielversprechende Kandidaten-Paare durch Kreuzungen generiert. Bei der visuellen Analyse dieser Doppelmutanten zeigte sich eine deutliche Zunahme der beobachtbaren Phänotypen gegenüber Einzelmutanten. Die Informationen zu den Doppelmutanten und die Liste homologer Gene sind über eine im Internet verfügbare Benutzerschnittstelle verfügbar, die im Rahmen dieser Arbeit entwickelt wurde.

Auf Basis einer großen Anzahl von Sequenzen, die aus der Bestätigung einzelner Insertionen in GABI-Kat stammen, konnten gemeinsame Charakteristika von T-DNA Insertionen aufgezeigt werden. Die Integration der T-DNA führt fast immer zu einer Veränderung der Ursprungssequenz in Form von kleineren Deletionen (ca. 1-50 bp) und kleineren Duplikationen (ca. 1-10 bp). In seltenen Fällen wurden auch Inversionen und große Deletionen von mehreren Kilobasen beobachtet. Am Übergang zwischen T-DNA und Kerngenom von *A. thaliana* finden sich normalerweise kürzere Bereiche von Mikrohomologie, die in Einzelfällen bis zu 20 bp groß sind, oder ein *Filler*, dessen Ursprung in den meisten Fällen im Genom von *A. thaliana* zu finden ist und der wiederum oft kurze Bereiche von Mikrohomologie zur jeweils benachbarten Sequenz aufweist. Des Weiteren ist die T-DNA häufig um einige Basen bezüglich der erwarteten Schnittstelle in der *Border* verkürzt. Alle diese Eigenschaften einer T-DNA Insertion deuten darauf hin, dass die Integration der T-DNA hauptsächlich mit Hilfe des Reparaturmechanismus *Non-Homologous End-Joining* (NHEJ) stattfindet. Aber auch die Beteiligung anderer Reparaturmechanismen wie *Microhomology-Mediated End-Joining* (MMEJ) und *Synthesis-Dependant Strand-Annealing* (SDSA) ist wahrscheinlich. Diese Annahmen konnten in dieser Arbeit erstmals mit einer großen Fallzahl untersucht und belegt werden.

Einführung

Seit der Sequenzierung des Genoms der Ackerschmalwand (*Arabidopsis thaliana*) im Jahr 2000 ist eine der Schlüsselaufgaben zum weiteren Verständnis des Genoms die Erforschung der Funktionen der darin kodierten Gene. Zu diesem Zweck sind Linien, in denen ein Defekt in einem Gen vorliegt (*Knockouts*), ein wichtiges Werkzeug. Verschiedene Projekte in der Pflanzenforschung haben deshalb versucht, das Genom von *A. thaliana* mit Mutanten für möglichst alle Gene abzudecken. Das am häufigsten zur Umsetzung verwendete Werkzeug ist die mittels *Agrobacterium tumefaciens* vermittelte Transformation. Dabei wird die Transfer-DNA (T-DNA) des Tumor-induzierenden (Ti)-Plasmids in das Wirts-Genom eingeschleust. An der Insertionsstelle können dadurch kodierende Bereiche unterbrochen und Genfunktionen aufgehoben werden. Nach Identifizierung der genauen Insertionsstelle sind zielgerichtete Arbeiten mit *Knockouts* in bestimmten Genen möglich, um Rückschlüsse vom Genotyp auf den Phänotyp ziehen zu können. Darauf aufbauende neue Fragestellungen befassen sich mit der möglichst genauen Vorhersage der Insertionspositionen innerhalb des Genoms oder mit Problemen bei der Arbeit mit sehr ähnlichen (paralogen) Bereichen, die molekulare Arbeiten erschweren und im Fall von duplizierten Genen keine eindeutigen Ergebnisse liefern. Die genauen Mechanismen und strukturellen Konsequenzen einer T-DNA Insertion sind bis heute nicht vollständig verstanden.

Einige wichtige Fakten zum Modellorganismus *A. thaliana*, dessen Genom im Verlauf der Evolution mehrere Duplikationsereignisse durchlaufen hat und daher viele paraloge Bereiche aufweist, werden im folgenden Kapitel aufgeführt. In verschiedenen Insertionslinienpopulationen wurde *A. thaliana* durch das Bakterium *A. tumefaciens* transformiert, um eine große Menge von *Knockout*-Mutanten zu erstellen. Daher wird *A. tumefaciens* und bekannte Fakten über den Integrationsmechanismus vorgestellt. Des Weiteren werden Details über DNA-Reparaturmechanismen für Doppelstrangbrüche behandelt, die offenbar eine wichtige Rolle bei der T-DNA Integration spielen. Darauf folgt eine Einführung in die Insertionslinien-Kollektion GABI-Kat sowie die zugrunde liegende Datenbank und ihre Benutzer-Schnittstellen, die die Datenbereitstellung, Analysen und Benutzeranfragen ermöglichen.

2.1 Modellorganismus *Arabidopsis thaliana*

Die Modellpflanze *Arabidopsis thaliana* (Acker-Schmalwand) stammt aus der Familie der *Brassicaceae*, die kultivierte Pflanzen wie Senf, Kohl und Rettich beinhaltet und wurde im 16. Jahrhundert durch Johannes Thal (daher *thaliana*) im Harz entdeckt. Sie ist hauptsächlich in Europa, Asien und Nordamerika zu finden. Je nach Herkunft lassen sich verschiedene Ökotypen unterscheiden. Einer der Meistbenutzten ist Col-0, der ursprünglich in Columbia (Missouri) isoliert wurde. Weitere wichtige Ökotypen sind Landsberg *erecta* (Ler-0) aus Landsberg in Deutschland und Wassilewskija (Ws) aus dem gleichnamigen Ort in Russland.

Die Blätter von *A. thaliana* formen eine Rosette an der Basis der Pflanze mit vereinzelten Blättern am Stängel, an dessen Haupt- und Seitentrieben sich die Blüten bilden. Der schnelle Lebenszyklus von nur sechs Wochen unter optimalen Bedingungen von der Keimung bis zu reifen Samen sowie die geringe Größe der Pflanze und ihres Genoms machen *A. thaliana* zu einem idealen Modellorganismus, der bereits seit den 40er Jahren in der Pflanzenforschung Anwendung findet [Laibach, 1943]. Sie ist außerdem die erste Pflanze deren Genom im Jahr 2000 sequenziert wurde [The Arabidopsis Genome Initiative, 2000].

2.1.1 Das Genom von *A. thaliana*

Seit die erste Genomsequenz von *A. thaliana* im Jahr 2000 veröffentlicht wurde [The Arabidopsis Genome Initiative, 2000], wurde stetig an der Verbesserung der Sequenz und der Annotation gearbeitet. Das Genom von *A. thaliana* ist ein verhältnismäßig kleines Pflanzengenom, das fünf Chromosomen umfasst. In dieser ersten Version des Genoms betrug die Genomgröße 115 Megabasen und es wurden 25.498 Protein-kodierende Gene annotiert (Gen-Dichte von 4.5 kb pro Gen). Als Grundlage dienten auf den Chromosomen verankerte *Bacterial Artificial Chromosome* (BAC)-Sequenzen. Bis in das Jahr 2003 fanden weitere Verbesserungen der Genomsequenz und -annotation durch *The Institute for Genomic Research* (TIGR) statt. Die aktuellste Genomversion stammt von *The Arabidopsis Information Resource* (TAIR) in der Version 10 aus dem Jahr 2010.

Genom-Annotation TIGR5 Im Jahr 2003 erschien die letzte auf BAC-Sequenzen basierende Genom-Annotation TIGR5. TIGR wurde mit dem Ziel der Re-Annotation des *A. thaliana* -Genoms gegründet und gefördert, da seit der Veröffentlichung des Genoms eine Vielzahl heterogener Annotationen entstanden war. Durch Einbindung weiterer Sequenzen wuchs die Größe des assemblierten Genoms auf 119 Megabasen. Zur weiteren Verbesserung der bestehenden Annotationen dienten cDNA- und *Expressed Sequence Tag* (EST)-Daten. Die Zahl der vorhergesagten Protein-kodierenden Gene betrug nach dieser Re-Annotation 27.384, was einer Dichte von einem Gen pro 4.4 kb entspricht [Wortman et al., 2003]. Eine einheitliche Gen-Nomenklatur wurde eingeführt. Im sogenannten AGI-Code wird

zunächst der Organismus-Name angegeben (At für *A. thaliana*), die Chromosomen-Nummer (bzw. P für das Plastom und C für das Chondriom), gefolgt von einem G für „Gen“ (oder beispielsweise TE für transponierbare Elemente) und einer fünfstelligen Nummer. So steht der Bezeichner *At4g23270* für ein Gen auf Chromosom 4 im Genom von *A. thaliana*, in diesem Fall eine Protein-Kinase.

Genom-Annotation TAIRv10 Nach TIGR fanden weitere Verbesserung der Annotation unter der Verantwortung von TAIR statt. Die aktuellste Genom-Annotation ist TAIR in der Version 10 (TAIRv10). Diese Annotation beruht auf Chromosomen. Gene sind somit nicht mehr mit ihren Positionen auf den BACs, sondern mit ihren Positionen auf dem jeweiligen Chromosom annotiert. Zusätzlich zu Protein-kodierenden Genen und Pseudogenen enthält diese Annotation auch RNA-Gene und Gene für transponierbare Elemente. Vorhandene Annotationen sowie alternative *Splice*-Varianten konnten insbesondere durch die Nutzung von RNA-Sequenzierungen verbessert werden. Die TAIRv10 Genom-Veröffentlichung enthält 27.416 Protein-kodierende Gene, 4.827 Pseudogene und transponierbare Elemente, sowie 1.359 nicht-kodierende RNAs bei einer Gesamt-Genomgröße von 119,2 Megabasen (30,4 Mbp für Chromosom 1, 19,7 Mbp für Chromosom 2, 23,5 Mbp für Chromosom 3, 18,6 Mbp für Chromosom 4 und 27 Mbp für Chromosom 5) [TAIRv10, 2010].

2.1.2 Genomevolution in *A. thaliana*

Neben verschiedenen Möglichkeiten zur Entstehung von Mutationen sind insbesondere Ereignisse, die zu duplizierten Bereichen im Genom führen, eine wichtige Triebkraft in der Evolution von Genomen. Aufgrund von Duplikationsereignissen in der Evolution von *A. thaliana* weist deren Genom viele paraloge Bereiche auf. Dies stellt ein Problem für die Arbeit mit dem Modellorganismus dar - beispielsweise wenn ein untersuchtes Gen in mehreren Kopien vorliegt und es nicht genügt eines von beiden auszuschalten, um eine Veränderung des Phänotyps zu beobachten. Dies ist auch der Grund dafür, dass eine eindeutige Zuordnung von kurzen Sequenzen zum Genom durch Sequenzvergleiche, sowie das Entwerfen von Amplifizierungs-Primern für eine PCR in diesen Bereichen schwierig ist. Derartige redundante Bereiche haben hauptsächlich drei Ursachen: Genomduplikationen, Tandem-Duplikationen und transponierbare Elemente [Krebs et al., 2014].

Genomduplikation Eine Genomduplikation entsteht durch Polyploidisierung, d.h. die Anzahl der Chromosomen wird vervielfacht. Zu unterscheiden sind Allopolyploidie und Autopolyploidie. Ursache für beide ist zunächst die Bildung von unreduzierten Gameten aufgrund eines Fehlers in der Reifeteilung (Meiose). Von Autopolyploidie spricht man, wenn dies innerhalb einer Art geschieht, von Allopolyploidie bei Kreuzung von zwei reproduktiv kompatiblen Spezies. Die entstehenden tetraploiden Nachkommen sind mit dem diploiden Ursprungsorganismus

kreuzbar, allerdings sind die daraus entstehenden triploiden Nachkommen aufgrund unpaarbarer Chromosomen während der Meiose steril.

Auf diese Weise gewonnene Chromosomen können wieder komplett oder teilweise (beispielsweise durch ungleiche Crossing-Overs) verloren gehen. Des Weiteren bieten gewonnene Gen-Kopien die Möglichkeit der Ansammlung von Mutationen. Denn selbst bei einer Mutation, die ein Gen nicht-funktional macht, ist sichergestellt, dass noch eine weitere funktionierende Kopie im Genom vorhanden ist. Meistens wird eine von beiden Kopien durch die Vielzahl an Mutationen defekt (Nonfunktionalisierung). Es ist aber auch möglich, dass eine Kopie eine neue Funktion übernimmt (Neofunktionalisierung) oder beide Kopien dieselbe Funktion teilen (Subfunktionalisierung). Duplikationen jeglicher Art ermöglichen die Entwicklung neuer Varianten eines Gens [Krebs et al., 2014].

In der Evolutionsgeschichte von *A. thaliana* gab es drei Haupt-Duplikationsereignisse (siehe Abbildung 2.1) [Tang et al., 2008]. Das erste Duplikationsereignis fand vor etwa 120 Millionen Jahren statt (γ), zwei weitere stammen aus jüngerer Zeit etwa 50 (β) bzw. 30 (α) Millionen Jahre in der Vergangenheit. Aufgrund des letzten Duplikationsereignisses haben 89 % der Gene in *A. thaliana* eine duplizierte Kopie. 51,6 % der duplizierten Gene gehen auf die vorletzte Duplikation zurück, während von der ersten Genomduplikation noch 20,3 % erhalten sind [Bowers et al., 2003]. Wie in Abbildung 2.2 zu sehen, sind die duplizierten Bereiche aufgrund von ungleichen Crossing-Overs (siehe nächster Absatz) mittlerweile sehr zufällig im Gesamt-Genom verteilt.

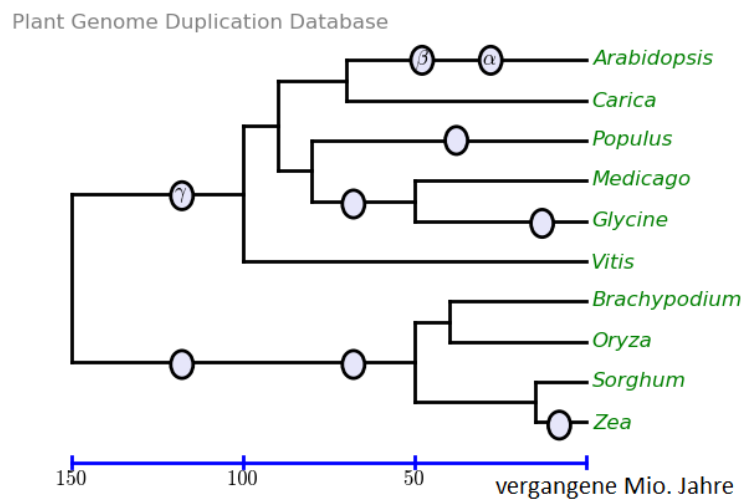


Abbildung 2.1: Zeitverlauf der Genomduplikationen von *A. thaliana* und verwandter Gattungen. Während der Evolution von *A. thaliana* hat es insgesamt drei Duplikationsereignisse (α , β , γ) gegeben, von denen zwei aus verhältnismäßig junger Zeit stammen. Abbildung von PGDD [2015].

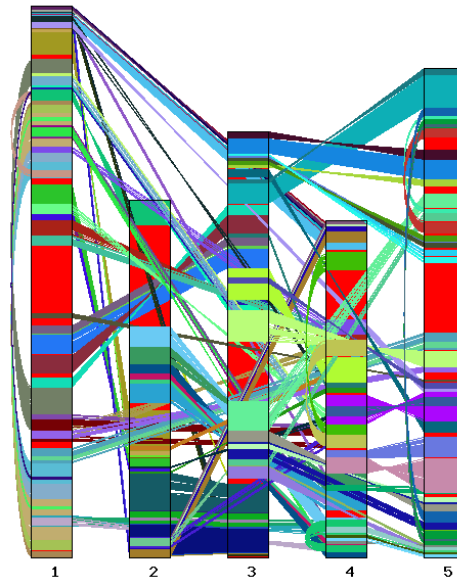


Abbildung 2.2: Paraloge Bereiche in den fünf Chromosomen von *A. thaliana*. Große Teile des Genoms weisen mindestens einen verwandten Bereich innerhalb des Genoms auf. Abbildung aus Haas et al. [2005].

Tandemduplikationen Tandemduplikationen sind eine weitere treibende Kraft der Evolution von Genomen. Normalerweise findet ein Crossing-Over während der Meiose zwischen identischen Sequenzen von zwei Chromosomen statt und Teile der Arme werden untereinander ausgetauscht. Die Größe der beiden beteiligten Chromosomen wird dadurch nicht verändert. Findet diese Paarung jedoch an zwei anderen homologen Sequenzen statt, spricht man von ungleichem Crossing-Over. Resultat eines solchen Ereignisses sind zwei verschieden große Chromosomen (siehe Abbildung 2.3). Ausgangspunkt für eine solche Paarung können verschiedene ähnliche Sequenzen sein. Insbesondere transponierbare Elemente können, durch ihre Ähnlichkeit zueinander, die für ein ungleiches Crossing-Over nötige Sequenzhomologie zur Verfügung stellen [Krebs et al., 2014].

Transponierbare Elemente Transponierbare Elemente (TE) sind genomische Abschnitte, die innerhalb des Genoms mobil sind. Grob unterscheiden lassen sich zwei Typen von transponierbaren Elementen: DNA-Transposons können mit Hilfe einer Transposase, die sie oft (aber nicht zwangsläufig) selbst enthalten und exprimieren, herausgeschnitten (oder kopiert) und an einem anderen Ort wieder integriert werden. Dabei entstehen aufgrund des versetzten Schnittes in der Zielregion und anschließender Reparatur Duplikationen an der Insertionsstelle. Nach erneutem Herausschneiden des Transposons bleiben diese duplizierten Bereiche als *Repeats* zurück. Bei den sogenannten Retrotransposons wird zunächst ein RNA-Intermediat als Kopie des transponierbaren Elements erstellt, das mittels einer reversen Transkriptase in DNA umgeschrieben und an einer anderen Stelle im Genom

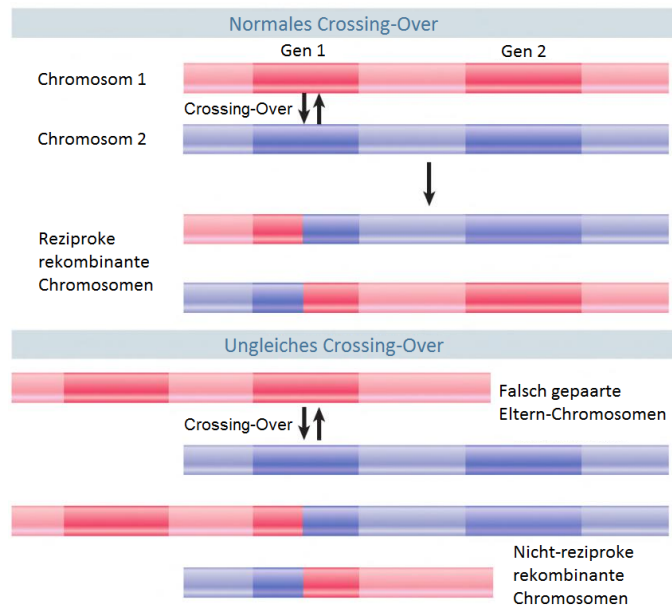


Abbildung 2.3: Crossing-Over Varianten. Bei einem normalen Crossing-Over werden Sequenzteile zwischen den beiden Chromosomen ausgetauscht. Die Größe der Chromosomen bleibt davon unberührt. Durch Paarung zweier nicht an derselben Position liegender homologer Bereiche findet ein ungleiches Crossing-Over statt und führt zu zwei ungleichen Chromosomen, von denen eines duplizierte Gene aufweist, die dem anderen fehlen. Abbildung aus Krebs et al. [2014].

integriert wird. Die meisten TE integrieren mehr oder weniger zufällig und oft sind viele Kopien im gesamten Genom verteilt. TE sind zum Einen selbst der Grund für ähnliche Bereiche im Genom, zum Anderen können sie auch aufgrund ihrer Sequenzähnlichkeit Ausgangspunkt für ein ungleiches Crossing-Over sein (siehe Abbildung 2.4), was sich auch in einem signifikant erhöhten Anteil von TE innerhalb von mutmaßlichen Duplikationen widerspiegelt [Hughes et al., 2003].

Als Resultat all dieser Ereignisse kodieren nur 9,7 % der etwa 27.000 Proteinkodierenden Gene in *A. thaliana* für ein einzigartiges Protein, alle anderen besitzen mindestens ein paraloges Gen [Armisen et al., 2008]. Teil von Gen-Familien mit mehr als 5 Mitgliedern sind 40 % der Gene, teilweise bestehen diese Familien aus mehr als 100 Mitgliedern [Cannon et al., 2004; Wortman et al., 2003].

Diese Zahlen verdeutlichen, dass Phänotyp-Analysen auf Basis eines einzelnen Gen-*Knockouts* in vielen Fällen nicht ausreichend sind und dass sehr ähnliche Regionen innerhalb des Genoms die molekularbiologische Arbeit sehr erschweren können. Insbesondere bei Sequenzvergleichen von kurzen Sequenzen mit dem Programm BLAST [Altschul et al., 1997] ist das Ergebnis aufgrund der paralogen Natur des Genoms von *A. thaliana* oft nicht eindeutig. In manchen Bereichen des Genoms ist es nicht möglich, eindeutige Primer für eine PCR zu generieren.

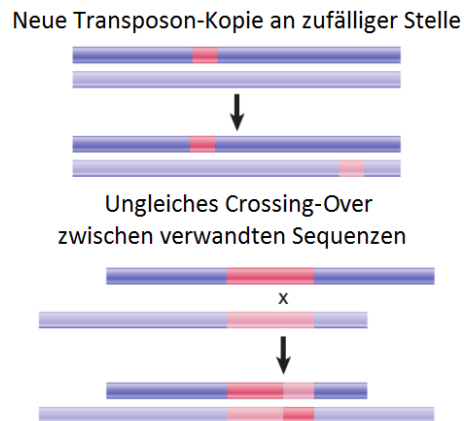


Abbildung 2.4: Ein Transposon führt zu ungleichem Crossing-Over. Ein zufällig integriertes Transposon kann durch seine Ähnlichkeit zu einer an einem anderen Ort integrierten Kopie ein ungleiches Crossing-Over bewirken. Abbildung aus Krebs et al. [2014].

2.1.3 Reverse Genetik und die GABI-Kat-Kollektion

Zur Aufklärung von Genfunktionen können generell zwei Verfahren unterschieden werden: Vorwärtsgenetik und reverse Genetik. Ersteres beschreibt dabei den klassischen Ansatz, einen Phänotyp zu charakterisieren und darauf aufbauend Gene zu bestimmen, die ihn verursachen. Die für einen bestimmten Phänotyp verantwortlichen Gene zu identifizieren ist keine leichte Aufgabe und erfordert oft die systematische Untersuchung vieler mutagenisierter Einzelorganismen. Exemplare, die den untersuchten Phänotyp zeigen, werden genauer untersucht, um mögliche Kandidaten-Gene zu identifizieren [Hartwell et al., 2008].

Bei der reversen Genetik wird der umgekehrte Weg gegangen. Ausgehend von einem ausgeschalteten Gen wird versucht, die Auswirkungen auf den Phänotyp zu ermitteln, um Rückschlüsse auf die Funktion des Gens zu ziehen. Das Mittel der Wahl sind *Knockout*-Mutanten, die eine Mutation in genau dem Gen zeigen, an dem Interesse besteht. In Bakterien lassen sich solche Mutationen relativ einfach zielgerichtet durch Ausnutzung der homologen Rekombination erstellen. In Pflanzen steht dieses Werkzeug allerdings nur eingeschränkt zur Verfügung und es kommen zufälliger Verfahren wie EMS-Mutagenese oder T-DNA Insertionen zur Anwendung. Die durch *A. tumefaciens* vermittelte Transformation bietet sich besonders an, da der Insertionsort anhand von einem *Flanking Sequence Tag* (FST) vorhergesagt werden kann, der basierend auf der bekannten T-DNA Sequenz durch PCRs und anschließender Sequenzierung erstellt wird (siehe Kapitel 2.4.2). Aufgrund der Zufälligkeit des Insertionsortes bedarf es einer Vielzahl von Mutanten, um den gewünschten *Knockout* zu erzielen. Durch die Erstellung einer großen Kollektion von *Knockout*-Mutanten, die interessierten Forschern zur Verfügung gestellt wird, lässt sich diese Zufälligkeit gut umgehen. Dies war das Ziel des GABI-Kat Projektes (siehe Kapitel 2.4).

2.2 Transformation durch *Agrobacterium tumefaciens*

Bakterien der Gattung *Agrobacterium* haben eine außergewöhnliche Eigenschaft gemeinsam. Sie sind in der Lage einen Teil ihrer Gene in Pflanzenzellen einzuschleusen. Dies führt normalerweise zu unkontrolliertem Tumorwachstum im infizierten Gewebe. Durch die eingeschleusten Gene stellen Zellen dieses Gewebes nach erfolgter Infektion dem Bakterium Stoffwechselprodukte zur Verfügung, die dieses metabolisieren kann [Schell et al., 1979].

Vertreter der Gattung sind beispielsweise *Agrobacterium rhizogenes*, das in infizierten Pflanzenteilen das Wachstum sogenannter Haarwurzeln auslöst [Intrieri and Buiatti, 2001] oder *Agrobacterium vitis*, der Verursacher der Wurzelhalsgallen genannten Tumore in Weinreben [Ophel and Kerr, 1990]. Der am besten untersuchte Vertreter dieser Gattung ist jedoch *Agrobacterium tumefaciens*. Seit Beginn des 20. Jahrhunderts ist bekannt, dass *A. tumefaciens* für Wurzelhalsgallen in einer Vielzahl von zweikeimblättrigen Pflanzen verantwortlich ist. Als Ursache dieser Tumorbildung wurde 1977 die Fähigkeit nachgewiesen, Gene eines Plasmids stabil in das Pflanzengenom zu integrieren [Chilton et al., 1977]. Diese Fähigkeit macht *A. tumefaciens* seitdem zu einem vielseitig einsetzbaren Werkzeug in der Pflanzenforschung.

2.2.1 *A. tumefaciens* und das Ti-Plasmid

A. tumefaciens (Familie *Rhizobiaceae*, Abteilung der *Proteobacteria*) ist ein gramnegatives Bodenbakterium [Smith and Townsend, 1907]. Es ist nicht darauf angewiesen, eine Pflanze zu infizieren, sondern ist selbstständig überlebensfähig. Neben einem zirkulären (ca. 2,8 Mbp) und einen linearen Chromosom (ca. 2,1 Mbp) umfasst dessen Genom zwei weitere Plasmide mit ungefähr 542 kb und 214 kb [Goodner et al., 2001]. Das kleinere Plasmid enthält die sogenannte Transfer-DNA (T-DNA) und wird als Ti-Plasmid (Ti = Tumor-induzierend) bezeichnet [Zambryski et al., 1983]. Nur ein kleiner Teil der Agrobakterien trägt allerdings ein Ti-Plasmid und ist deshalb als Pflanzenpathogen einzustufen. Neben einer oder mehrerer Kopien der T-DNA, die den Abschnitt definiert, der in das Pflanzengenom integriert wird, enthalten alle bekannten Varianten dieses Plasmids eine Virulenz(*vir*)-Region. Sie enthält viele für die Integration wichtige Gene.

Die T-DNA Region wird von zwei 25 bp langen imperfekten *Repeats* flankiert (*Left* und *Right Border* - LB/RB) [Tinland, 1996; Duerrenberger et al., 1989]. Des Weiteren enthält die T-DNA Onkogene, die für Proteine zur Synthese von Phytohormonen kodieren, mit denen die Tumor-induzierenden Eigenschaften assoziiert werden [Van Larebeke et al., 1974]. Weitere auf der T-DNA kodierte Proteine sind für die Biosynthese von Opinen verantwortlich. Diese Stoffe können wiederum von *A. tumefaciens* metabolisiert werden [Schell et al., 1979]. Die Gene für den Opinkatabolismus sind auf dem Teil des Ti-Plasmids kodiert, der nicht in das Pflanzen-

genom integriert wird.

Durch Ausschalten der Pathogenität (Entfernen der Onkogene aus der T-DNA) und Austauschen des Bereiches zwischen den flankierenden Bereichen der T-DNA durch nahezu beliebige Sequenzen können Pflanzen gezielt transformiert werden [Hernalsteens et al., 1980]. Da dies im Gegensatz beispielsweise zur Genkanone [Sanford et al., 1987] eine recht elegante und verlässliche Methode zur Transformation von Pflanzen darstellt entwickelte sich *A. tumefaciens* zu einem wichtigen Werkzeug. Heutzutage wird ein binäres Vektorsystem verwendet, bei dem die *vir*-Gene auf einem separaten Plasmid enthalten sind. Das Plasmid, das die T-DNA enthält, kann dadurch klein gehalten werden, um Klonierungsarbeiten zu erleichtern [Bevan, 1984].

Neben dem Einbringen neuer Gene, was insbesondere in der Entwicklung von transgenen Nutzpflanzen von Interesse ist, ist eine weitere nützliche Eigenschaft einer T-DNA Insertion, dass ein an der Insertionsstelle vorhandenes Gen aufgrund der Größe der integrierten T-DNA in der Regel nicht mehr funktionsfähig ist. Dies konnte genutzt werden, um große Kollektionen von transformierten Linien mit *Knockout*-Mutanten zu erstellen. Details sind in Kapitel 2.4 beschrieben.

2.2.2 Der Weg der T-DNA in den Zellkern

Für eine Integration der T-DNA in das Genom der Pflanze muss die T-DNA zunächst zwei Hindernisse überwinden: Die Zellwand und Zellmembran der Pflanze sowie die Doppelmembran des Zellkerns, der die genomische DNA vom Zytoplasma trennt. Eine schematische Darstellung dieses Prozesses, bei dem mehrere Vir-Proteine (VirA, VirB, VirD1, VirD2, VirD4, VirD5, VirE2, VirE3, VirF und VirG) eine wichtige Rolle spielen, ist in Abbildung 2.5 dargestellt.

Nur zwei der *vir*-Gene sind ständig exprimiert: *virA* und *virG*. Die von ihnen kodierten Proteine bilden zusammen ein Zwei-Komponenten Regulationssystem [Albright et al., 1989]. VirA bildet ein die Innenmembran von *A. tumefaciens* durchspannendes Homodimer, das phenolische Stoffe wie Acetosyringon (das von verwundeten Pflanzenzellen abgesondert wird) erkennt [Lee et al., 1995]. Binden diese Erkennungs-Stoffe an VirA führt dies zur Phosphorylierung des Transkriptionsfaktors VirG durch VirA. VirG bindet dann an die 12 bp lange *vir*-Box, aufwärts des *vir*-Operons, wodurch die Transkription der übrigen *vir*-Gene eingeleitet wird, die normalerweise nicht transkribiert werden [Jin et al., 1990; Gelvin, 2003].

Daraufhin wird eine einzelsträngige Kopie der T-DNA erstellt, der sogenannte T-Strang. Die Helikase VirD1 und die Endonuklease VirD2 arbeiten dabei zusammen als Endonuklease-Komplex [Tinland et al., 1994]. Die Schnittstelle befindet sich genau vor der Erkennungssequenz 5'-CAGGATATATT-3' [Jasper et al., 1994]. VirD2 bindet anschließend an die *Right Border* (RB) und lenkt den T-Strang in die Pflanzenzelle durch ein aus VirB und VirD4 bestehendes Typ-IV-Sekretionssystem [Gelvin, 2010]. Des Weiteren schützt VirD2 die RB vor Zersetzung durch Exonukleasen in der Pflanzenzelle [Duerrenberger et al., 1989].

Neben dem T-Strang und VirD2 gelangen die Virulenz-Proteine VirD5, VirE2,

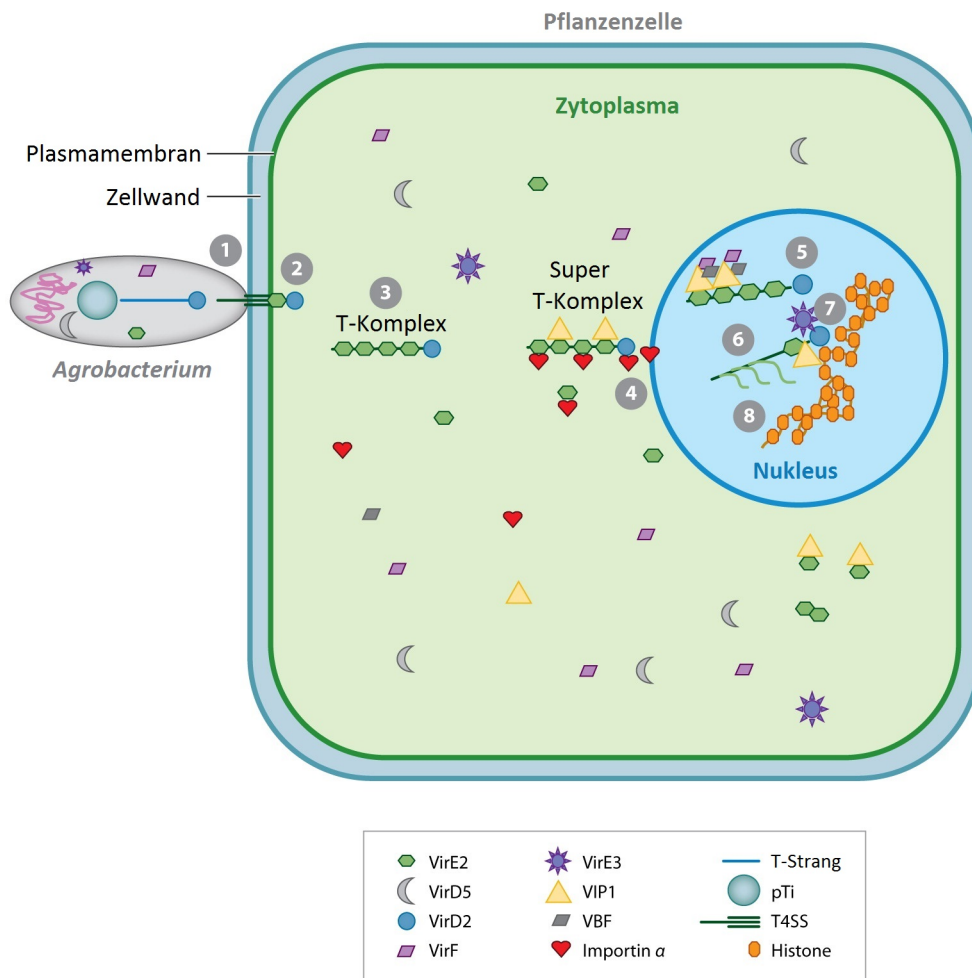


Abbildung 2.5: Der Weg der T-DNA in die Pflanzenzelle und in den Zellkern. Nach Andocken von *A. tumefaciens* an die Pflanzenzelle (1) wird ein T-Strang durch die Plasmamembran mittels eines Typ-IV-Sekretionssystems und VirD2 in die Zelle geschleust (2). Durch Anlagerung von VirE2 entsteht dort der T-Komplex (3), der in den Kern transportiert wird (4). Der T-Komplex lokalisiert das Chromatin (5), die schützende Proteinhülle wird entfernt (6), die T-DNA wird integriert (7) und darauf kodierte Gene exprimiert (8). Abbildung aus Gelvin [2010].

VirE3 und VirF durch ein C-terminales Export-Signal ebenfalls durch das Typ-IV-Sekretionssystem in das Zytoplasma der Pflanzenzelle [Vergunst et al., 2005]. VirE2-Proteine bedecken dabei den kompletten T-Strang und schützen ihn vor Degradierung [Citovsky et al., 1989]. VirE3 bindet an VirE2 und vermittelt durch ein Kern-Lokalisations-Signal den Transport dieses T-Komplexes in den Nukleus [Lacroix et al., 2005].

Wie in Abbildung 2.5 angedeutet, gab es Hinweise, dass das Pflanzenprotein VIP1

(VirE2 interacting protein) ebenfalls eine Rolle bei der Kernlokalisierung von VirE2 und bei der Interaktion mit Histonen (was auf eine Beteiligung bei der Integration hindeuten würde) spielt [Tzfira et al., 2001, 2002; Li et al., 2005]. Neuere Ergebnissen widersprechen jedoch dieser Annahme [Shi et al., 2014]. VirF unterstützt bei der Degradierung des schützenden Proteinkomplexes [Tzfira et al., 2004b; Schrammeijer et al., 2001] und wird dabei wiederum selbst durch VirD5 vor einer Degradierung durch das Ubiquitin-Proteasom-System der Pflanzenzelle geschützt [Magori and Citovsky, 2011]. Nach Degradierung der schützenden Proteinhülle wird der einzelsträngige T-Strang wahrscheinlich in doppelsträngige (ds) T-DNA umgewandelt [Tzfira et al., 2004a]. Ein Modell dazu wurde 2013 publiziert [Liang and Tzfira, 2013] und ist in Abbildung 2.6 dargestellt.

Dieses Modell erklärt neben der Bildung von ds T-DNA auch, wie sogenannte T-Ringe entstehen können. T-Ringe sind T-DNA-Moleküle, die sich ähnlich wie ein Plasmid zu einem Ring schließen. Sie können zwar nicht mehr in das Genom der infizierten Zelle integriert werden, aber darauf vorhandene Gene können transient exprimiert werden [Rolloos et al., 2014; Singer et al., 2012].

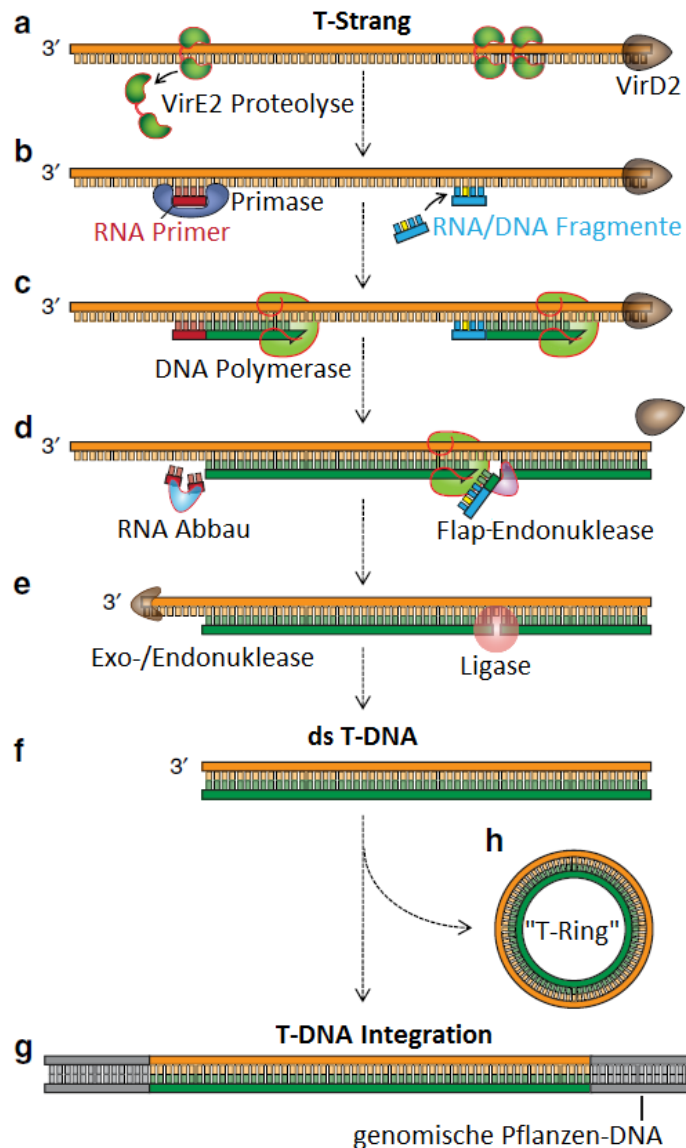


Abbildung 2.6: Modell zur Bildung von ds T-DNA innerhalb des Zellkerns der Pflanzenzelle. Die an den T-Strang gebundenen VirE2-Proteine werden mittels Proteolyse entfernt (a). Ein RNA-Primer wird durch die Primase erzeugt oder existierende DNA/RNA-Fragmente werden als Primer genutzt. Dabei kann es zu Fehlpaarungen (in Gelb) kommen (b). Der komplementäre Strang wird durch Verlängerung dieser Primer mit Hilfe einer DNA-Polymerase in 5'-3'-Richtung synthetisiert (c) und die Primer schließlich degradiert (d). Das immer noch einzelsträngige 3'-Ende der T-DNA ist anfällig für einen Abbau durch Endo- und Exonukleasen (e,f). Die ds T-DNA integriert in das Pflanzengenom (g) oder bildet T-Ringe (h). Abbildung aus Liang and Tzfira [2013].

2.2.3 Integration in das Genom

Während die Proteine, die am Transport des T-Stranges in den Zellkern der Pflanze gut untersucht sind, ist über die an der Integration selbst beteiligten Proteine weniger bekannt. Von den exportierten Vir-Proteinen scheint keines eine Rolle bei der Integration zu spielen. *In vitro* konnte gezeigt werden, dass VirD2 die Ligation von zwei T-DNA *Borders* katalysieren kann [Pansegrau et al., 1993]. Des Weiteren kann die vor Exonukleasen schützende Funktion von VirD2 einen indirekten Einfluss auf das Resultat der Integration haben: eine zu großen Teilen intakte RB.

Im Jahr 2004 wurden die vorherrschenden T-DNA Integrationsmodelle in einem Review zusammengefasst [Tzfira et al., 2004a]. Eine Übersicht über die dort vorgestellten verbreitetsten Modelle zeigt Abbildung 2.7. Diese drei Modelle stellen unterschiedliche Möglichkeiten zur Integration der T-DNA dar und werden nachfolgend beschrieben.

Doppelstrangbruch-Reparatur Modell Voraussetzung für die Integration in diesem Modell ist ein Doppelstrangbruch in der genomischen Wirts-DNA und die Überführung des einzelsträngigen T-Stranges in eine doppelsträngige T-DNA. Entwundene Enden oder durch Exonukleasen entstehende Überhänge binden an offene Enden des Doppelstrangbruches und werden schließlich repariert und ligiert. Ausreichend für die Bindung sind kurze zueinander homologe Sequenzen von wenigen Basen, sogenannte Mikrohomologien. Überstehende Enden werden durch Endo- oder Exonukleasen entfernt.

Einzelstrangbruch-Reparatur Modell In diesem Modell integriert die T-DNA in Form eines einzelsträngigen T-Stranges. Die Tatsache, dass die T-DNA als einzelsträngiges Molekül in den Nukleus gelangt, ist das stärkste Argument für diesen Mechanismus. Ausgangspunkt ist ein Einzelstrangbruch, der durch eine Exonuklease zu einer Lücke vergrößert wird. Die einzelsträngige T-DNA bindet wiederum an Mikrohomologien. Überhängende Enden werden durch eine Endonuklease entfernt und der T-Strang mit der genomischen DNA ligiert. Anschließend wird der nicht passende Gegenstrang durch Endo- und Exonukleasen entfernt und daraufhin mit der integrierten T-DNA als Vorlage komplementiert.

Mikrohomologie-abhängiges Modell Das Mikrohomologie-abhängige Modell ist eine Erweiterung des Einzelstrangbruch-Reparatur Modells. Es basiert auf der Annahme, dass VirD2 als Ligase fungiert [Tinland and Hohn, 1994], was jedoch kontrovers ist [Ziemienowicz et al., 2000]. T-DNA Integration beginnt in diesem Modell mit Mikrohomologie-basierter Bindung des 3'-Endes des T-Stranges mit der LB an einen entwundenen Strang der Ziel-DNA. Nach Entfernung des überhängenden Endes und einem Schnitt der Ziel-DNA bindet auch das andere Ende des T-Stranges an einen Teil mit Mikrohomologie im Gegenstrang und wird mit Hilfe der Ligase-Funktion von VirD2 an das offene Ende der Ziel-DNA ligiert. Der Gegenstrang wird wie im Einzelstrangbruch-Reparatur Modell repariert.

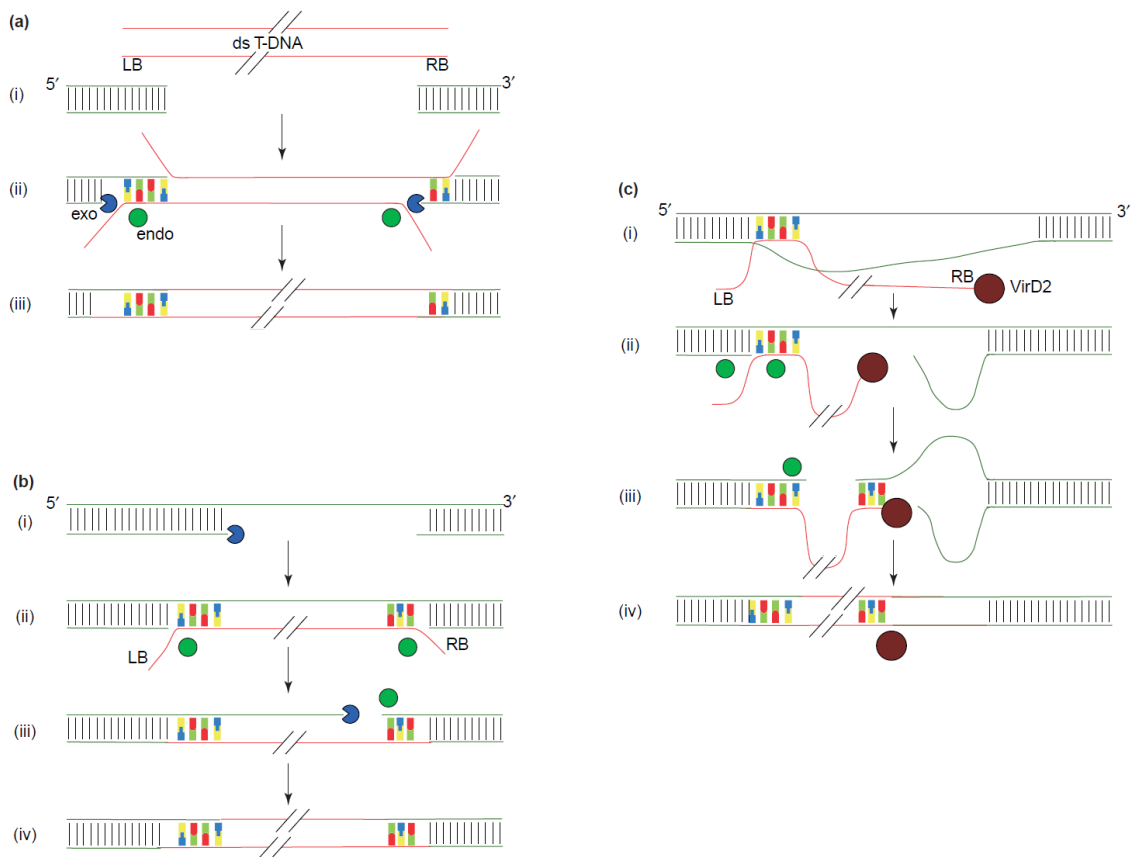


Abbildung 2.7: Modelle der T-DNA Integration. Im Doppelstrangbruch-Reparatur Modell (a) integriert die T-DNA als doppelsträngiges Molekül (i), Endo- und Exonukleasen reparieren die Bruchstellen (ii-iii). Im Einzelstrangbruch-Reparatur Modell (b) integriert die T-DNA als einzelsträngiges Molekül in eine Lücke (i). Überhänge und der intakte Strang werden durch Endo- und Exonukleasen entfernt (ii-iii) und die T-DNA komplementiert (iv). Im Mikrohomologie-abhängigen Modell (c) beginnt die Integration anhand einer Bindung durch Mikrohomologie (i), gefolgt von einem Schnitt der genomischen DNA durch eine Endonuklease (ii), Ligierung durch VirD2 (iii) und Komplementierung wie im Einzelstrangbruch-Reparatur Modell (iv). Abbildung aus Tzfira et al. [2004a].

Wahrscheinlich sind Mechanismen des Wirts an der Integration beteiligt, um z.B. die Nuklease-Funktionalität zur Verfügung zu stellen. Diese Theorie wird durch eine verringerte Transformations-Effizienz bzw. die Verhinderung von Transformation in verschiedenen *Knockout*-Mutanten von *A. thaliana* gestärkt [Nam et al., 1998, 1999; Mysore et al., 2000; Anand et al., 2007]. Im wahrscheinlichsten Modell integriert die T-DNA als doppelsträngiges Molekül in Doppelstrangbrüche [Salomon and Puchta, 1998; Chilton and Que, 2003; Tzfira et al., 2003]. Dabei sind Proteine

des Reparatur-Mechanismus der Pflanze involviert [van Attikum et al., 2001].

Auch in Hefen kann eine Integration über die Ausnutzung von Reparatur-Mechanismen stattfinden, wobei hauptsächlich die homologe Reparatur (siehe Kapitel 2.3.1) zum Einsatz kommt [van Attikum and Hooykaas, 2003], während nicht-homologe Reparatur (Kapitel 2.3.2) der für die Integration von T-DNA in Pflanzen dominante Mechanismus zu sein scheint [Ray and Langer, 2002; Britt and May, 2003].

Für eine Integration als doppelsträngiges Molekül in einen Doppelstrangbruch spricht auch, dass die Abwärtsregulierung der Expression des für XRCC4 kodierenden Gens (einem der Schlüsselproteine für nicht-homologe Reparatur) die Anzahl stabiler Transformationen in *A. thaliana* und *Nicotiana benthamiana* (Tabak) erhöht, während Hochregulierung der Genexpression desselben Gens die T-DNA Integrationseffizienz verringert. Diese Ergebnisse lassen sich dadurch erklären, dass im Fall der Abwärtsregulierung die Reparatur von auftretenden Doppelstrangbrüchen verzögert wird, wodurch T-DNA Integration in diese offenen Brüche erleichtert wird. Aufgrund der Interaktion von VirE2 und XRCC4 wird vermutet, dass VirE2 die Aktivität von XRCC4 unterdrücken kann und zu einer verzögerten Reparatur beiträgt [Vaghchhipawala et al., 2012].

Ein weiteres Argument für ein Modell in dem die T-DNA als doppelsträngiges Molekül integriert, liegt in der Vielfältigkeit der beobachteten Insertionsmuster. Bei vielen T-DNA Insertionen wurde festgestellt, dass die T-DNA in mehreren fusionierten Kopien vorliegt. Dabei wurden alle möglichen Kombinationen von Fusionen bezogen auf den Übergang zur ursprünglichen pflanzlichen DNA beobachtet (LB-RB, RB-LB, LB-LB, RB-RB). Bestimmte Kombinationen sind nur möglich, wenn eine Rekombination über die RB stattfindet. Dies erfordert ein doppelsträngiges T-DNA Molekül [Krizkova and Hroudá, 1998; De Buck et al., 1999].

2.2.4 Ort der T-DNA Insertion

Durch Auswertung der Insertionsorte in großen Kollektionen von T-DNA Insertionslinien konnte untersucht werden, ob der Insertionsort im Genom zufällig ist. Ein auffällig hoher Teil der Insertionen ist in Regionen mit höherer Gendichte zu finden [Alonso et al., 2003]. Zudem scheint eine Neigung zur Integration in Transkriptionsstartpunkte sowie poly(A)-Regionen zu existieren [Szabados et al., 2002; Li et al., 2006]. Dies deutet zunächst auf einen nicht zufälligen Integrationsmechanismus hin, allerdings können diese Beobachtungen durch eine Selektionsverzerrung erklärt werden. Durch die Selektion anhand einer durch die T-DNA vermittelten Resistenz (siehe Kapitel 2.4.1) werden Insertionen innerhalb von transkriptionell aktiven Bereichen des Genoms bevorzugt. In inaktiven Bereichen ist es sehr wahrscheinlich, dass auch das übertragene Resistenz-Gen inaktiv ist. Es ist daher anzunehmen, dass der genaue Ort der Insertion im Genom zufällig ist, was auch der Schluss einer Studie zur Untersuchung von T-DNA Insertionen unter nicht-selektiven Bedingungen ist [Kim and Gelvin, 2007].

2.3 Reparaturmechanismen für Doppelstrangbrüche

Die Integration der T-DNA in einen Doppelstrangbruch unter Ausnutzung der in *A. thaliana* vorkommenden Reparaturmechanismen scheint, wie im vorigen Kapitel beschrieben, der wahrscheinlichste Integrationsmechanismus zu sein. Ein Bruch beider DNA-Stränge stellt eine große Gefahr für die Stabilität des Genoms dar. Unrepariert kann er zu Chromosomenbruch und Zelltod, falsch repariert zu Chromosomen-Translokation und Tumoren führen [Hoeijmakers, 2001]. Die Ursachen für Doppelstrangbrüche sind beispielsweise ionisierende Strahlung, oxidative freie Radikale oder Endonukleasen [Yoshiyama et al., 2013]. Zellen besitzen verschiedene Mechanismen, um offene Enden eines DNA-Doppelstrangbruches zu reparieren. Das Resultat der Reparatur hängt stark vom verwendeten Reparaturmechanismus ab und führt zu verschiedenen Veränderungen der Ausgangssequenz.

In Hefen und Bakterien ist homologe Reparatur (HR) am häufigsten vertreten, während in höheren Organismen nicht-homologe Reparatur durch *Non-Homologous End-Joining* (NHEJ) der dominante Mechanismus ist [Wyman and Kanaar, 2006]. Neben homologer und nicht-homologer Reparatur kann die Reparatur auch durch Ausnutzung von kurzen Homologie-Abschnitten geschehen. Dieser *Microhomology-Mediated End-Joining* (MMEJ) genannte Mechanismus zeichnet sich durch seine Unabhängigkeit von Schlüsselproteinen für HR und NHEJ aus.

2.3.1 Homologe Reparatur

Voraussetzung für die Reparatur durch homologe Rekombination ist das Vorhandensein einer homologen Vorlage, die zur Reparatur verwendet werden kann [Wyman et al., 2004]. Eine Übersicht über den homologen Reparaturmechanismus zeigt Abbildung 2.8. Dabei kann die homologe Vorlage durch einen Mechanismus der *Synthesis-Dependant Strand-Annealing* (SDSA) genannt wird, genutzt werden, um ein Ende des offenen Doppelstrangbruches neu zu synthetisieren, bevor die Lücke über andere Mechanismen geschlossen wird.

Lagern sich beide offene Enden des Doppelstrangbruches an die homologe Vorlage an, bildet sich eine *Holliday*-Struktur, in der beide Stränge anhand der Vorlage neu synthetisiert werden. Die Auflösung dieser Struktur führt entweder zu einer Trennung der beiden DNA-Stränge oder zu einem Austausch zwischen beiden (Crossing-Over) [Wyman and Kanaar, 2006].

Eine entscheidende Rolle bei der Reparatur durch homologe Rekombination spielen die Proteine Rad51 und Rad52. Rad52 bindet einzelsträngige DNA-Moleküle und kann die DNA-DNA-Interaktion komplementärer DNA-Stränge vermitteln [Symington, 2002]. Des Weiteren interagiert es mit der Rekombinase Rad51, die ein helikales Nukleoproteinfilament auf einzelsträngiger DNA bildet und homologe Paarung mit doppelsträngiger DNA ermöglicht [Galkin et al., 2006].

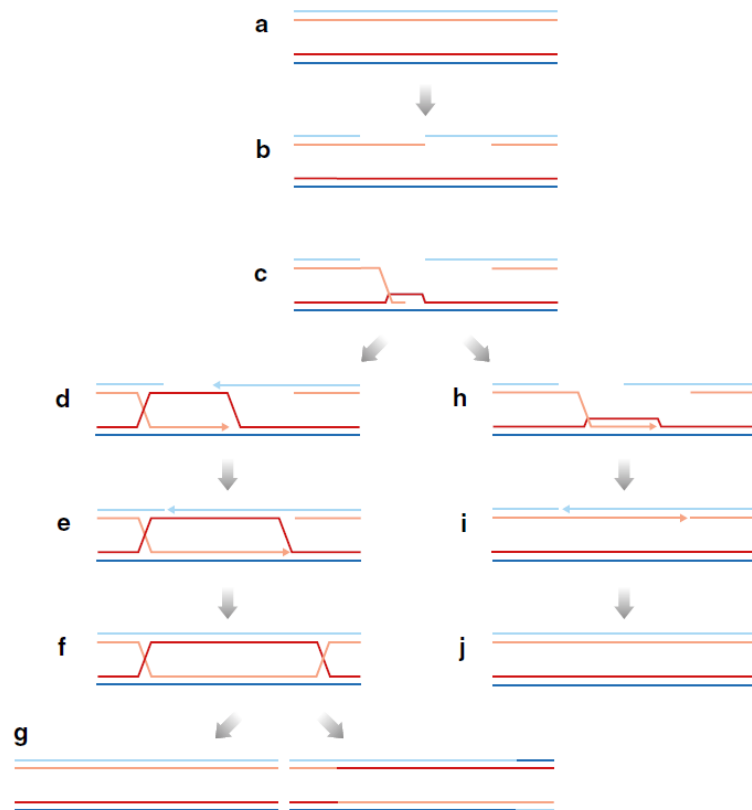


Abbildung 2.8: Reparatur durch homologe Rekombination. Es entsteht ein Doppelstrangbruch mit einzelsträngigen 3'-Enden (a,b). Diese bilden das Substrat für eine durch Rekombinase vermittelte Nukleoproteinfilament-Formation, die zur Homologie-Erkennung dient und einen DNA-Strang Austausch zwischen den beiden homologen Doppelsträngen ermöglicht (c). Dies resultiert in einer D-Loop genannten Struktur in dem der nicht paarende Strang der Vorlage mit dem anderen Strang der Doppelstrangbruch-DNA paart (d). Nach DNA-Synthese (e), Verbindung der offenen Enden und Bildung einer *Holliday*-Struktur (f), kommt es entweder zu einem Crossing-Over zwischen beiden Molekülen oder zu einer Trennung (g). Als Alternative zur Bildung und Auflösung einer *Holliday*-Struktur kann durch den SDSA-Mechanismus die Synthese in der homologen Vorlage auch nur durch Paarung eines einzelnen Stranges erfolgen und der andere Strang des Doppelstrangbruchs anhand der neu synthetisierten Vorlage komplementiert werden (h,i,j). Grafik aus Wyman and Kanaar [2006].

Ebenfalls verwandt mit Reparatur durch homologe Rekombination ist *Single-Strand-Annealing* (SSA), das ebenfalls Rad52 erfordert und sich besonders zur Reparatur von repetitiven Bereichen eignet. Dabei findet die Bindung der beiden offenen Enden an längeren Bereiche von mehr als 25 bp Homologie statt, gefolgt von

der Reparatur des Doppelstrangbruches [Symington, 2002; Sugawara et al., 2000]. Charakteristisch für diesen Reparatur-Mechanismus sind große Deletionen des Bereiches zwischen den beiden repetitiven Sequenzen.

Abgesehen von großen Deletionen durch SSA ist als Resultat einer Reparatur durch HR die Reparaturstelle meistens unverändert bezüglich der Ausgangs-Sequenz. Ein Crossing-Over kann allerdings zu einem Austausch eines homologen Teils führen, was insbesondere zur Transformation von Bakterien genutzt werden kann. SDSA kann durch das Eindringen in nicht-homologe Bereiche auch zu Sequenzen mit anderem Ursprung an der Reparaturstelle führen. Die Bindung zur Vorlage ist in solchen Fällen schwach und wird oft nach kurzer Synthese wieder abgebrochen. Dadurch entstehen an der Reparaturstelle sogenannte *Filler* mit Sequenzen aus einem anderen Teil des Genoms. *Filler* sind kurze Sequenzstücke, die keinem der beiden Enden des Doppelstrangbruches zugeordnet werden können. SDSA kann durch Unterbrechung der Synthese und Wechsel der Vorlage zu komplexeren *Fillern*, bestehend aus unterschiedlichen Sequenzteilen, führen [Gorbunova and Levy, 1999].

2.3.2 Nicht-homologe Reparatur

In Pflanzen spielt Homologe Reparatur (HR) nur eine untergeordnete Rolle, hauptsächlich bei der Reparatur von auftretenden Brüchen während der Synthese-Phase der Mitose. Doppelstrangbrüche werden aus Ermangelung einer nahe der Bruchstelle vorhandenen Vorlage hauptsächlich mittels NHEJ repariert (siehe Abbildung 2.9). Dies hat in der Regel Deletionen von mehreren Basenpaaren, kleine Duplikationen oder *Filler* an der Reparaturstelle zur Folge und ist damit fehlerbehafteter als homologe Reparatur. NHEJ ist abhängig von den Proteinen Ku70 und Ku80, die als Heterodimer an offene Enden binden und weitere Proteine wie XRCC4 oder DNA-Ligase IV rekrutieren [Wyman and Kanaar, 2006].

Ein weiteres Merkmal von NHEJ ist, dass eine kurze Mikrohomologie zwischen den beiden zu verbindenden Enden von bis zu 5 bp erforderlich scheint. Allerdings können offene Enden auch mittels einfacher Ligation verbunden werden, was keine Veränderung der Sequenz zur Folge hat. Dies erfordert jedoch intakte Enden. *Filler*, die manchmal während der Reparatur mittels NHEJ auftreten, können durch SDSA erklärt werden (siehe Kapitel 2.3.1). Merkmale einer Reparatur durch NHEJ ist eine, im Gegensatz zur homologen Reparatur, oft veränderte Reparaturstelle [Gorbunova and Levy, 1999].

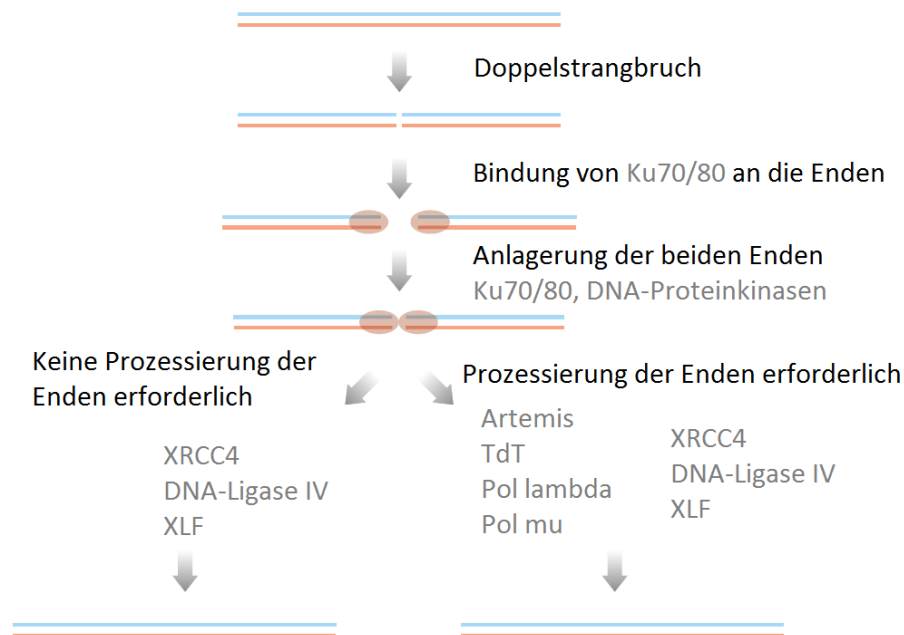


Abbildung 2.9: Reparatur durch NHEJ. Die offenen Enden eines Doppelstrangbrüches werden durch den Ku70/Ku80 Heterodimer gebunden, durch den DNA-Proteinkinasen die beiden offenen Enden zusammenbringen. XRCC4, DNA-Ligase IV und XLF sind die verantwortlichen Proteine für die Verbindung des Doppelstrangbrüches. Wenn eine Prozessierung der Enden nötig ist, um ligierbare Enden zu produzieren, kann das durch die Nuklease Artemis oder die Polymerasen TdT, pol- λ oder pol- μ geschehen. Grafik aus Wyman and Kanaar [2006].

2.3.3 Mikrohomologie-abhängige Reparatur

Neben homologer Reparatur und NHEJ gibt es einen dritten, schlechter charakterisierten Mechanismus zur Reparatur von Doppelstrangbrüchen. Obwohl bereits seit fast 30 Jahren bekannt ist, dass es Reparatur-Mechanismen unabhängig von homologer Rekombination gibt, die mehr als 5 bp Mikrohomologie benutzen [Roth and Wilson, 1986], hat dieser Mechanismus erst seit relativ kurzer Zeit Aufmerksamkeit erhalten [McVey and Lee, 2008]. MMEJ benutzt größere Bereiche von Mikrohomologie von 6-25 bp und funktioniert dabei unabhängig von Ku70/Ku80 und DNA-Ligase IV (siehe Abbildung 2.10). Ku80-defiziente Mutanten weisen größere Bereiche von Mikrohomologie an den reparierten Stellen in Hefe und Maus auf [Boulton and Jackson, 1996; Liang et al., 1996]. Bei einer Mikrohomologie-Größe von 6-8 bp scheint MMEJ auch weitgehend unabhängig von Rad52 zu sein, was darauf hindeutet, dass es sich um einen eigenständigen Mechanismus handelt. Ab einer Mikrohomologie-Größe von 8 bp ist Rad52 zunehmend involviert, was einen fließenden Übergang von MMEJ zu SSA (siehe Kapitel 2.3.1) nahelegt [Daley and Wilson,

2005]. Eine Reihe von Proteinen scheint mit MMEJ assoziiert zu sein [McVey and Lee, 2008]. Eine entscheidende Rolle spielt dabei offensichtlich der MRX-Komplex, bestehend aus Mre11, Rad50 und Xrs2 (bzw. NBS1 in Säugetieren). Dieser sorgt für die Teilentfernung eines Stranges, um die zur Reparatur nötige Mikrohomologie freizulegen. Weiterhin wird die Rad1-Rad10 Endonuklease zur Entfernung überhängender Enden nach erfolgter Reparatur benötigt.

Charakteristisch für eine Reparatur mittels MMEJ sind Deletionen von variierender Größe und die Insertion von einzelnen Nukleotiden an der Reparaturstelle, was durch die Fehleranfälligkeit der verwendeten Polymerase erklärt werden kann [McVey and Lee, 2008].

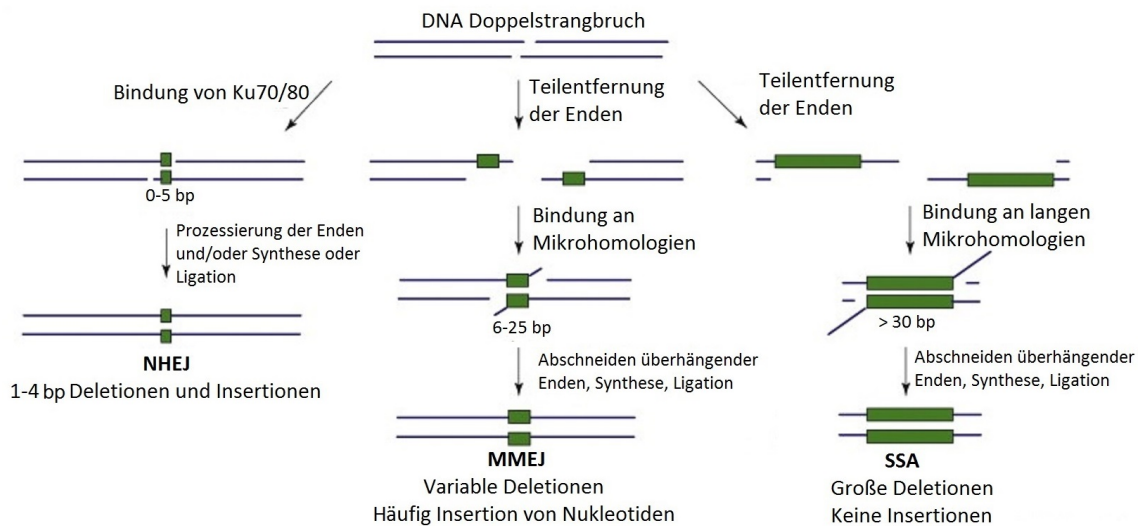


Abbildung 2.10: Einordnung von MMEJ zwischen NHEJ und SSA. Bei der Reparatur durch NHEJ verhindert das Ku70/Ku80-Heterodimer durch seine Bindung an ein offenes Ende die Resektion der offenen Enden. Durch Bindung an kurze Mikrohomologie, Auffüllung entstandener Lücken durch die DNA-Polymerase sowie Ligation durch DNA-Ligase IV wird der Bruch repariert. Entfällt die Bindung durch Ku70/Ku80, werden Teile der offenen Enden entfernt und gegebenenfalls etwas größere Homologie-Bereiche freigelegt, die mittels MMEJ oder SSA repariert werden. In beiden Fällen müssen überhängende 3'-Enden vor Synthese und Ligation entfernt werden. MMEJ kann im Unterschied zu SSA zu einzelnen eingefügten Basen führen, während beide Mechanismen größere Deletionen als NHEJ zur Folge haben können. Abbildung aus McVey and Lee [2008].

2.4 Die Insertionslinienpopulation GABI-Kat

Das Ziel des GABI-Kat Projektes war die systematische Erstellung einer Sequenz-indizierten Kollektion von T-DNA *Knockout*-Mutanten in *A. thaliana*. Sequenz-indiziert bedeutet, dass über die T-DNA flankierende Sequenzen schnell eine Linie mit Insertion im gewünschten Bereich des Genoms identifiziert werden kann. Diese Information wird für alle generierten Linien zur Verfügung gestellt. Das Projekt startete 1999 im Rahmen des GABI (Genomanalyse am biologischen System Pflanze) Förderprogramms am Max-Planck-Institut in Köln (Kat = Kölner *Arabidopsis* T-DNA-Linien) und läuft seit Januar 2007 an der Universität Bielefeld. Die GABI-Kat-Kollektion ist weltweit die zweitgrößte öffentlich verfügbare Ressource für T-DNA Insertionslinien im Col-0 Ökotyp, übertroffen nur von der SALK-Kollektion in La Jolla (Kalifornien, USA).

2.4.1 Transformation in GABI-Kat

Zur Transformation der Linien in GABI-Kat wurde die *Floral dip*-Methode [Clough and Bent, 1998] angewandt. Dabei werden die Blüten von *A. thaliana* Pflanzen in eine Lösung getaucht, die mit dem gewünschten Vektor transformierte Agrobakterien enthält, sowie 10 % Saccharose und 0,02 % des Netzmittels Silwet L-77 [Rosso et al., 2003].

Der in GABI-Kat am häufigsten zur Transformation verwendete Vektor ist pAC161 (siehe Abbildung 2.11). Weitere Vektoren, die sich nur geringfügig davon unterscheiden sind pAC106, pADIS1 und pGABI1. Alle GABI-Kat Vektoren enthalten ein Gen, das die Resistenz gegen das Herbizid Sulfadiazin (Sul) vermittelt [Guerineau et al., 1990]. In pAC161 ist dieses Gen unter Expression eines starken konstitutiven 1'-2' Promotors aus *A. tumefaciens* [Velten et al., 1984]. Da die Vektoren ursprünglich für *Activation Tagging* (eine Methode bei der im Gegensatz zu *Knockouts* das Ziel ist, Gene nahe der Insertion durch Überexpression zu untersuchen) entworfen

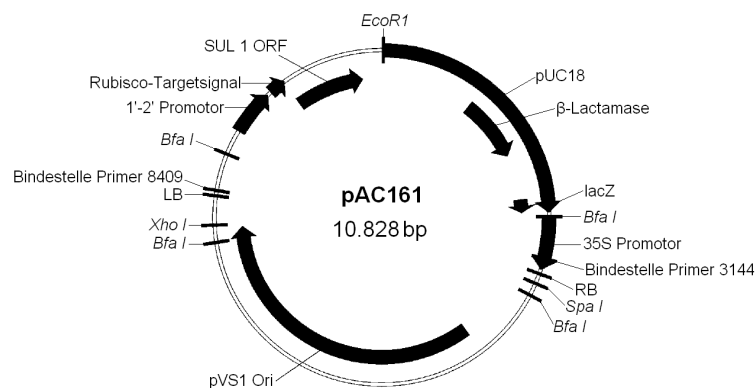


Abbildung 2.11: Vektor pAC161 mit enthaltener T-DNA. Grafik aus Rosso et al. [2003].

wurden, enthalten sie einen 35S-Promotor aus dem Blumenkohl-Mosaikvirus nahe der RB. Dieser kann bei einer T-DNA Insertion strangaufwärts eines Gens dazu führen, dass dieses stärker exprimiert wird. Die Schnittstellen in den beiden *Borders* liegen so, dass die *Left Border* (LB) der T-DNA um 3 bp gekürzt wird und 22 bp an der RB verloren gehen.

Die reifen Samen der transformierten Pflanzen wurden geerntet (T1) und auf Sul-Resistenz selektiert. Aus geerntetem Blattmaterial der daraus gewachsenen T1-Pflanzen wurde DNA extrahiert und genutzt, um *Flanking Sequence Tags* (FSTs) zu generieren (siehe Kapitel 2.4.2) und daraus die mutmaßliche Insertionsposition abzuleiten. Alle Linien wurden in 96er-Blöcken aufgezogen und die daraus entstehende Nomenklatur beibehalten. So setzt sich eine GABI-Kat Linien-ID aus der Block-Nummer und der Position im Block zusammen. Die Linie 456C06 stammt beispielsweise aus Block 456 an Position C06.

2.4.2 Insertionsstellen-Vorhersage

Die Integration der T-DNA innerhalb des *A. thaliana*-Genoms erfolgt weitgehend zufällig (siehe Kapitel 2.2.4). Um den Ort der Insertion zu bestimmen, können FSTs generiert werden. Ein FST ist eine Sequenz, die ausgehend von der T-DNA bis in den genomischen Bereich der Wirts-DNA reicht und die T-DNA „flankiert“. Durch Sequenzvergleiche kann anhand dieser ein potentieller Insertionsort bestimmt werden [Rosso et al., 2003; Strizhov et al., 2003].

FST-Generierung

Einen Überblick über die FST-Generierung in GABI-Kat gibt Abbildung 2.12. Zunächst wird die aus den T1-Pflanzen extrahierte DNA mit dem Restriktionsenzym *BfaI* verdaut. Es schneidet in der relativ kurzen Erkennungssequenz 5'-C|TAT-3'. Wie in Abbildung 2.11 zu sehen, ist sowohl in der Nähe der LB als auch der RB eine *BfaI*-Schnittstelle zu finden. Durch das zufällige Auftreten weiterer Schnittstellen im genomischen Teil nahe der Insertionsstelle entstehen ein oder mehrere Fragmente, die neben einem Teil der T-DNA auch einen Teil genomischer DNA enthalten. An die Schnittstellen wird ein Adapter ligiert. Darauf folgt eine lineare PCR mit dem innerhalb der T-DNA liegenden Primer 8474 (5'-ATAATAACGCTGCGGACATCTACATTTT-3') gefolgt von einer PCR mit exponentieller Amplifikation mit dem etwas näher an der LB liegenden Primer 8409 (5'-ATATTGACCATCATACTCATTGC-3') und einem Adapterprimer. Der Großteil der FSTs wurde an der LB generiert, ein kleinerer Teil an der RB, unter Verwendung der Primer 3144 (5'-GTGGATTGATGTGATATCTCC-3') und CR3S (5'-TTGAGCATATAAGAAACCCTTAGT-3'). Die Sanger-Sequenzierung des aus den PCRs entstandenen Produkts erfolgte mit dem 8409 Primer. Das Qualitäts- und Vektor-*Trimming* wurde mittels PHRED [Ewing and Green, 1998] und Pre-gap4 [Staden, 1996] durchgeführt.

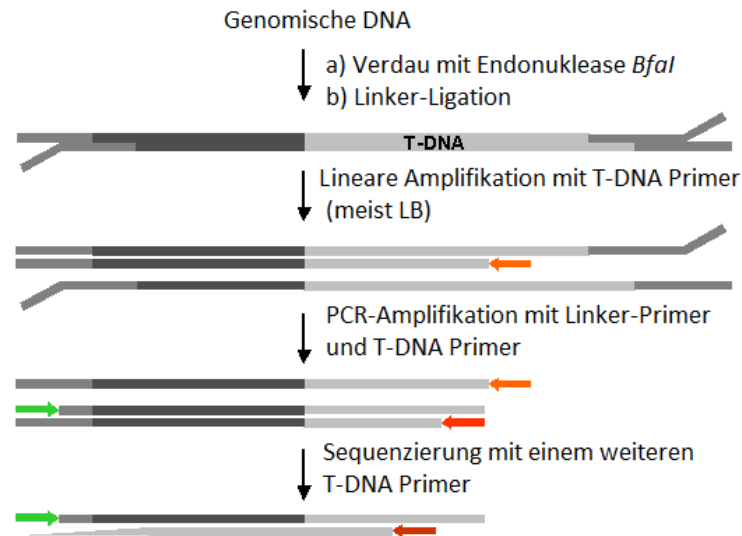


Abbildung 2.12: Übersicht über die FST-Generierung. Nach *BfaI*-Verdau und Ligation von Linkern an die Schnittstellen folgen zwei PCRs mit immer weiter am Rand der T-DNA liegenden Primern zur Anreicherung der Sequenz nahe der Insertion. Das entstehende Produkt wird schließlich mit einem weiter am Rand der T-DNA liegenden Primer sequenziert. Abbildung von der GABI-Kat Webseite [2015].

Vorhersage der Insertionsposition

Zur Bestimmung der mutmaßlichen Insertionsposition wurde ein Sequenzvergleich mittels BLASTn [Altschul et al., 1997] gegen die BAC-basierte TIGR *A. thaliana* Genomsequenz in der Version 5 (siehe Kapitel 2.1.1) durchgeführt [Li et al., 2006]. Alle Treffer, die unter einem Schwellenwert (e -Value kleiner als $5e^{-4}$) lagen, wurden gespeichert, und der Beginn des BLAST-Treffers in der Referenzsequenz wird als potentielle Insertionsposition auf dem jeweiligen BAC vermerkt. Liegt der Treffer in einem Gen, wird er als Gentreffer klassifiziert, ansonsten als Genomtreffer im intergenischen Bereich.

Die vielversprechendsten Treffer sind diejenigen die zwischen Transkriptionsstart und -ende liegen, wobei nicht zwischen Treffern in Exons und Introns unterschieden wird, da die T-DNA groß genug ist, um auch bei einer Insertion in einem Intron die Transkription zu unterbrechen. Diese Treffer sind sogenannte CDSi-Hits (CDS+Introns). Des Weiteren wurde ein Bereich von 300 bp um die CDSi-Region als 5'- bzw. 3'-Treffer definiert, da auch ein Treffer in der jeweiligen *untranslated region* (UTR) einen *Knockout* des betroffenen Gens bewirken kann. Berücksichtigt wurden Protein-kodierende Gene sowie Pseudogene.

2.4.3 Der Bestätigungs-Prozess (*Confirmation*)

Die mit einer vorhergesagten Insertionsposition und betroffenen Genen annotierten FSTs sind über das Webinterface „SimpleSearch“ (<http://www.gabi-kat.de/db/genehits.php>) zugänglich und Benutzer können darin nach Insertionen in Genen, die für sie von Interesse sind, suchen und eine Bestellung für die jeweilige Linie eingeben. Zur Zeit (Stand Februar 2015) sind insgesamt 133.150 FSTs über das Webinterface SimpleSearch (siehe Kapitel 2.4.4) zugänglich. Verfügbar sind dadurch 72.036 Linien. Von den 27.206 Protein-kodierenden Genen im Kerngenom von *A. thaliana* gibt es für 13.375 (49,17 %) eine vorhergesagte Insertion in der GABI-Kat-Kollektion.

Nach Eingang einer Bestellung für eine Insertion in einer bestimmten Linie wird versucht, die auf Basis von FSTs vorhergesagte Insertion zu validieren. Dieser *Confirmation* genannte Vorgang besteht im Wesentlichen daraus, die angefragte Linie aufzuziehen (Kapitel 2.4.3), eine oder mehrere PCRs mit Primern spezifisch für die jeweilige Insertion durchzuführen (Kapitel 2.4.3), sowie die Linien mit bestätigten Insertionen an Benutzer und das Samenzentrum Nottingham Arabidopsis Stock Centre (NASC) abzugeben (Kapitel 2.4.3) [Li et al., 2007]. Eine Übersicht über den gesamten Bestätigungs-Prozess in GABI-Kat zeigt Abbildung 2.13.

Aufzucht der Linien

Um die angefragten Linien zu bearbeiten, werden zunächst T2-Pflanzen auf Sul-Selektionsmedium angezogen. Anhand der Anzahl überlebender Pflanzen dieser Segregation lässt sich abschätzen, wie viele Insertionen die Linie insgesamt enthält. Bei über 85 % resistenten Keimlingen ist mehr als eine Insertion wahrscheinlich. Lässt die Segregation auf nur eine Insertion schließen, werden 12 Pflanzen im Gewächshaus angezogen, andernfalls 18, um die Chance zu erhöhen, eine Pflanze zu erhalten, die die gesuchte Insertion enthält. DNA der überlebenden Pflanzen wird extrahiert und für eine klärende „*Confirmation*-PCR“ verwendet.

Sollte eine Linie sehr schlecht keimen oder sollten keine überlebenden T3-Pflanzen vorhanden sein, was manchmal aufgrund einer defekten Sul-Resistenz vorkommt, werden diese noch einmal ohne Selektionsmedium angezogen.

Primerdesign und *Confirmation*-Sequenzen

Für die *Confirmation*-PCR wird ein Primer spezifisch für die untersuchte T-DNA *Border* (festgelegt durch den dazugehörigen FST) sowie ein „Gen-spezifischer“ Primer generiert, der im Bereich für die vorhergesagte Insertion bindet. Um diesen Primer zu ermitteln wird das Programm primer3 verwendet [Untergasser et al., 2012]. Dabei wird der Bereich 220-580 bp (RB) bzw. 270-620 bp (LB) von der potenziellen Insertionsposition entfernt als Zielregion benutzt. Das Amplikon dieser PCR wird mit beiden Primern sequenziert (siehe Abbildung 2.14) und die beiden Sequenzen (eine „*Border*-spezifische“ sowie eine „Gen-spezifische“ *Confirmation*-Sequenz)

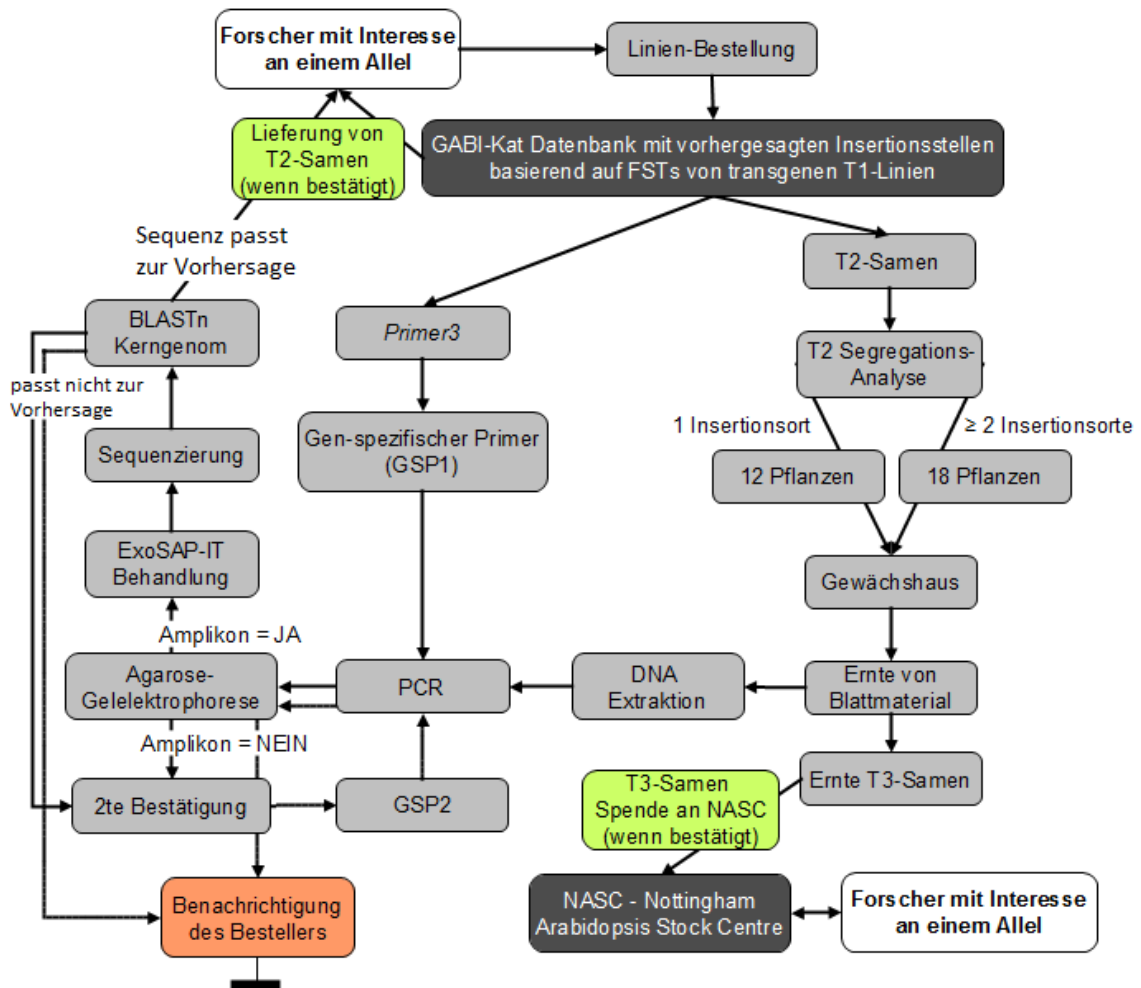


Abbildung 2.13: Übersicht über den GABI-Kat Bestätigungs-Prozess. Ein Forscher hat Interesse an einem bestimmten Allel in GABI-Kat. Nach eingegangener Bestellung folgt eine Segregation und Aufzucht der Linien im Gewächshaus. Mit aus Blattmaterial extrahierter DNA wird eine *Confirmation*-PCR unter Verwendung eines Gen-spezifischen Primers durchgeführt und bei Misslingen ein zweites Mal mit einem anderen Primer wiederholt. Bei erfolgreicher PCR wird das Produkt sequenziert und mittels BLASTn mit der Vorhersage abgeglichen. Bei Übereinstimmung werden dem Nutzer die T2-Samen geschickt und die geernteten T3-Samen an das NASC abgegeben. Abbildung aus Stracke et al. [2010]

werden mittels PHRED [Ewing and Green, 1998] und Pregap4 prozessiert [Staden, 1996]. Als Amplikon wird das reale Produkt einer PCR im Unterschied zum Amplimer bezeichnet, dass das theoretisch bei einer PCR entstehende Produkt benennt. Wenn ein darauf folgendes BLASTn der *Confirmation*-Sequenz gegen die *A. thaliana* Genomsequenz mit bis zu 500 bp Abweichung dasselbe Ergebnis liefert

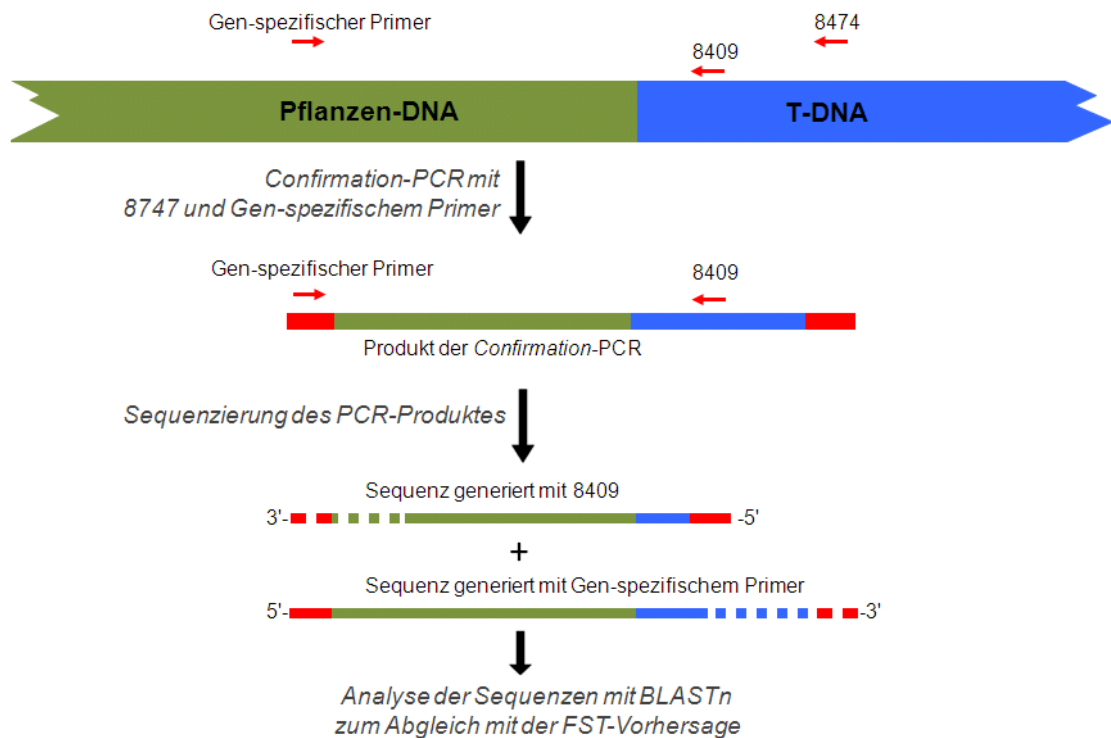


Abbildung 2.14: *Confirmation-PCR* und Sequenzierung einer Insertion an der LB. Zunächst wird eine PCR mit einem Gen-spezifischen Primer und dem *Border*-spezifischen Primer 8474 durchgeführt. Die Sequenzierung des Produkts geschieht mit dem etwas weiter am Rand der T-DNA liegenden Primer 8409 sowie mit dem Gen-spezifischen Primer. Die beiden Sequenzen werden mit der Genomsequenz abgeglichen. Bei der Gen-spezifischen Sequenz kann diese dabei etwas weiter in den Bereich der T-DNA reichen. Abbildung von der GABI-Kat Webseite [2015].

wie mittels FST vorhergesagt, gilt die Insertion als bestätigt. Zur Kontrolle der Anwesenheit der T-DNA im Genom wird zusätzlich eine PCR mit Primern durchgeführt, die spezifisch im Sul-Resistenz-Gen binden. Die dazu verwendeten Primer sind Sul2 (5'-GTGGAACCTTCAAAGCTGAAGT-3') sowie Sul4 (5'-ATTCACACAGGAAACAGCTATGA-3'). Schlägt eine erste *Confirmation-PCR* trotz erfolgreicher Sul-Kontrolle fehl, wird eine weitere PCR mit einem anderen Gen-spezifischen Primer durchgeführt, bevor keine weiteren Bestätigungsversuche mehr unternommen werden. Diese Insertion ist fehlgeschlagen und die entsprechende Linie kann nur noch von Benutzern bestellt werden, wenn sie weitere vorhergesagte Insertionen enthält.

Abgabe der Linien an Benutzer und Samenzentrum

Ist eine Insertion in einer Linie bestätigt, werden T2-Samen der jeweiligen Linie an den Benutzer verschickt. Zusätzlich werden T3-Samen an das NASC abgegeben. Einmal abgegebene Linien sind daraufhin nicht mehr über das Webinterface SimpleSearch bestellbar, auch wenn noch eine weitere Insertion vorhergesagt ist. Diese Linien können nur noch über NASC bezogen werden. Zum Zeitpunkt (Februar 2015) sind seit Beginn des Projekts 13.442 Linien zur weiteren Verteilung abgegeben worden.

2.4.4 GABI-Kat LIMS und SimpleSearch

Um die in GABI-Kat anfallenden Daten zu speichern und für die Öffentlichkeit zur Verfügung zu stellen, wurde eine Datenbank angelegt. Zugriff darauf erfolgt innerhalb der Universität Bielefeld hauptsächlich über das Labor Informations Management System (LIMS). Teile der zugrunde liegenden Datenbank sind über die Benutzer-Schnittstelle SimpleSearch im Internet verfügbar (siehe Abbildung 2.15).

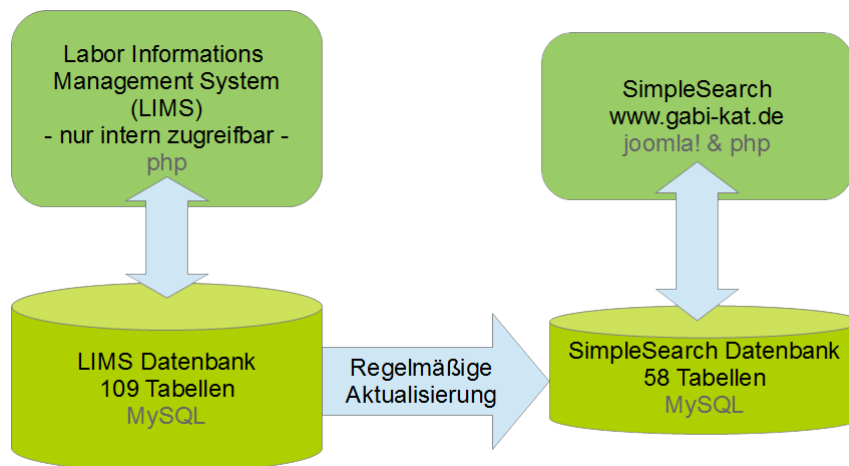


Abbildung 2.15: Struktur der GABI-Kat Datenbank und Weboberflächen. Auf die interne GABI-Kat Datenbank ist ausschließlich über die LIMS-Weboberfläche zugänglich. Die SimpleSearch-Datenbank enthält nur einen Teil der LIMS-Daten, der regelmäßig aktualisiert wird und über SimpleSearch allen Nutzern zur Verfügung steht.

GABI-Kat Datenbank Die GABI-Kat Datenbank basiert auf dem relationalen Datenbankverwaltungssystem MySQL. Das grundlegende Konzept ist eine Sammlung von Tabellen, die über bestimmte Spalten („Schlüssel“) miteinander verknüpft werden können, um größere Tabellen zu erhalten. Die Aufteilung von Daten auf mehrere Tabellen dient dabei der Vermeidung von redundanter Speicherung.

GABI-Kat LIMS Zur Koordinierung der Laborarbeiten wurde ein LIMS entwickelt. Es bietet Zugriff auf die GABI-Kat Datenbank und bereitet die darin gespeicherten Informationen in Listenform auf. Diese Listen enthalten die nötigen Details zu verschiedenen anfallenden Arbeiten, wie beispielsweise eine Liste aller PCRs mit Angaben zum Lagerort der zu verwendenden Primer und DNA. Des Weiteren bietet das LIMS diverse Möglichkeiten zur Dateneingabe, -manipulation und -analyse.

SimpleSearch Die Benutzeroberfläche SimpleSearch ist über die GABI-Kat Webseite unter www.gabi-kat.de zugänglich. Unter dem Menü-Unterpunkt „SimpleSearch“ gelangt der Benutzer zu einer Eingabemaske, mit der auf verschiedenen Wegen nach Insertionen in bestimmten Genen gesucht werden kann (siehe Abbildung 2.16).

The screenshot shows the SimpleSearch interface with the following content:

SimpleSearch

lines with genome hits: **72036** / genes with hits: **21382** / CDSi hits (protein-coding genes only): **13375**
64.17% or **21382** of all **33323** *A. thaliana* TAIRv10 genes are hit (chromosomal gene hits counted)
14.73% or **190** of all **1290** *A. thaliana* TAIRv10 ncRNA genes are hit (chromosomal gene hits counted)
71.45% or **19437** of the **27206** *A. thaliana* TAIRv10 protein coding genes are hit (chromosomal gene hits counted)
49.17% or **13375** of the **27206** *A. thaliana* TAIRv10 protein coding genes are hit (chromosomal CDSi hits counted)
lines at NASC as T3-sets: **13442**, **9128** of these contain at least one confirmed CDSi allele
56.48% or **7553** of the **13375** genes with chrom. CDSi hit predictions are confirmed for at least one allele and available from NASC
SimpleSearch / database content: GK27/20140513

Search for gene hits

Search by: gene code annotation text

enter one or more gene code e.g. 'At4g23270 ...' or a keyword e.g. 'kinase'

- 1** Search for gene hits
- 2** Search for line ID or GenBank ID of FST
- 3** Search for genome hits by BLAST
- 4** Search for genome range

Abbildung 2.16: SimpleSearch Benutzeroberfläche. Über die Oberfläche ist die Suche nach Insertionen anhand von vier verschiedenen Optionen möglich: anhand des AGI-Codes oder Annotationstextes (1), der Linie oder der Accession-Nummer eines FSTs (2), mittels BLAST einer benutzerdefinierten Sequenz (3) oder anhand einer Position auf den Pseudochromosomen (4). Screenshot von der GABI-Kat Webseite [2015].

Anhand dieser Suche gelangt der Benutzer zu Linien, die eine Insertion entsprechend seiner Suchkriterien enthalten. Abbildung 2.17 zeigt ein Muster für eine solche Linie. Neben Details zu FSTs sind außerdem Informationen zu verwendeten Primern, sowie *Confirmation*-Sequenzen in SimpleSearch verfügbar. Zusätzlich

SimpleSearch - Line and FST details

Line specific information

Line ID	051G10
Vector Used	pAC161
Line Availability	only available from NASC (N404882)
Segregation Analysis	50:35:27
Confirmed for Hit	At4g23270
Parent of DUPLO pair	none
Parent of pair(s)	5831, 10436, 10441, 75433, 75457, 75546, 75562, 75651, 75678, 75685

Gene hit **At4g23270**

```

>79-K012139-022-051-G10-8409
TTGATCCATGTAGATTNTCCCGGACATGAANGCCTTTACATGGATCCACTGCCATAATTC
TTATTTTTTTTAGTTGATTGAAACGTATTACGGTCTATTGATGAAAAAAGTGGCAAGA
GCAATACTTACTAAGTTCCAACACTCTTCTTGTGCATGGCTTCTGTTTGGTCCATCCCAA
ATATTCTCGCCATTCAAAATCTGCAATTTTAGGATTCATATCATCATCAAGAGGATGT
TCCCTGCTTTGAGGTCTCGATGTATAATTGTGAGTCGTGAATCTTGATGAAGATAAAGAA
TTCCTCTATGGGATCCTCCCTATAGTGAGTC

```

Sequence (A. th genome BLAST matches underlined)	
GenBank Accession	AL753564 [GenBank]
Graphic View	▼
Predicted Position of Insertion	Chr4:12173256 - go to primer design
BLAST e Value	2e-120
Hit Clone Code (BAC ID)	F21P8
Hit Gene Code	At4g23270 [TAIR] [MIPS] [SIGNAL]
Gene Annotation	cysteine-rich RLK (RECEPTOR-like protein kinase) 19
Insertion Classification	CDSi
Confirmation Status	confirmed, show confirmation sequences
Primer and wt-amplicons	show primer details

Abbildung 2.17: Details zu Linie 051G10. Im oberen Teil wird die Verfügbarkeit der Linie angezeigt. In diesem Fall ist die Insertion in *At4g23270* bereits bestätigt und die Linie deshalb nur von NASC erhältlich. Informationen über die Segregation werden angezeigt (ausgesäte Samen / gekeimte Samen / resistente Keimlinge), die Hinweise darauf geben können, wie viele Insertionen insgesamt in der Linie zu erwarten sind. In der gezeigten Linie 051G10 sind 27 von 35 Keimlingen resistent (77 %), was auf eine einzelne Insertion schließen lässt. Im unteren Teil werden FSTs sowie darauf basierende Insertions-Vorhersagen gelistet. Sind bereits Informationen zum Bestätigungsvorgang der jeweiligen Insertion verfügbar, werden diese hier nochmals angezeigt. Über weitere Links können *Confirmation*-Sequenzen sowie Details zu den Primern, die zur Bestätigung verwendet wurden, eingesehen werden. Screenshot von der GABI-Kat Webseite [2015].

gibt es eine Visualisierung eines Bereiches der Genomsequenz mit allen annotierten Genen und vorhergesagten sowie bestätigten Insertionen, die die Suche nach Insertionen in einem bestimmten Bereich des Genoms vereinfacht. Wurde durch den Benutzer eine interessante Insertion identifiziert und ist diese noch nicht bearbeitet (also nicht fehlgeschlagen oder bereits bestätigt und beim NASC verfügbar), kann er eine Bestellung über das entsprechende Webformular absetzen woraufhin die Bestellung in die GABI-Kat Datenbank eingegeben und wie in Kapitel 2.4.3 beschrieben bearbeitet wird.

2.4.5 Andere Kollektionen

Neben GABI-Kat gibt es noch eine Reihe weiterer Kollektionen, darunter mit SALK die weltweit größte Kollektion für T-DNA Insertionslinien in *A. thaliana*. Da in einer einzelnen Kollektion nicht die Insertionsallele für alle Gene enthalten sind, ergänzen sich diese Kollektionen gegenseitig. Drei davon, die für Analysen in dieser Arbeit von Bedeutung waren, werden im Folgenden vorgestellt.

SALK-Linien Das SALK T-DNA Projekt startete im Jahr 2001. Im Laufe des Projekts wurden 150.000 transgene Linien in Col-0 mit einer angenommenen Zahl von etwa 225.000 Insertionen generiert [Alonso et al., 2003]. Als Vektor wurde pROK2 verwendet. Im Unterschied zu GABI-Kat Vektoren enthält dieser keine Sul-Resistenz sondern eine Kanamycin-Resistenz als Selektionsmarker. Die Linien werden über das Arabidopsis Biological Resource Center (ABRC) zur Verfügung gestellt. Eine Bestätigung der einzelnen Insertionen findet nicht statt.

SAIL-Linien Die „Syngenta Arabidopsis Insertion Library“ (SAIL) besteht aus etwa 54.000 Linien. Transformiert wurden sie mit den Vektoren pCSA110 oder pD-AP101, die beide eine BASTA-Resistenz (gegen das Herbizid Glufosinat) tragen. FSTs wurden mittels eines modifizierten „*thermal asymmetric interlaced*“ (TAIL)-PCR Protokolls generiert. In einer TAIL-PCR werden Bereiche, die eine bekannte Region (in diesem Fall die T-DNA) flankieren in drei aufeinander folgenden Reaktionen mit verschachtelten Primern und degenerierten Primern aufkonzentriert und anschließend sequenziert. Typischerweise wird eine Reihe verschiedener degenerierter Primer benutzt, um die Wahrscheinlichkeit zu maximieren, ein flankierendes Produkt zu erhalten [Liu et al., 1995]. Die etwa 54.000 Linien in der SAIL-Kollektion enthalten etwa 85.000 Insertionen, die mittels BLAST ermittelt wurden. 30 % davon liegen in transkribierten Bereichen, 44 % in Promotoren und 26 % in intergenischen Regionen [Sessions et al., 2002]. Alle SAIL-Linien sind über das ABRC bestellbar, eine Bestätigung der Insertion ist dem Besteller selbst überlassen.

Wisc-Linien Die Kollektion der Universität Wisconsin enthält ca. 60.000 Linien. Im Unterschied zu GABI-Kat, SALK und SAIL ist hier der Wassilewskija (Ws)

Ökotyp zur Transformation mit dem Vektor pD991 verwendet worden. Dieser Vektor enthält eine Kanamycin-Resistenz. Um Insertionen in Genen von Interesse zu finden, wurden mehrere PCRs mit gepoolter DNA von 2.025 Einzel-Linien durchgeführt. In einer zweiten Runde wurde im Pool, der die Insertion enthält eine weitere Reihe PCRs von Pools mit 225 Linien gemacht, um die Linienzahl weiter einzuzugrenzen. Der Benutzer erhält 25 Samenröhrchen mit je 9 Linien aus denen er die Linie, die den *Knockout* enthält, über weitere PCRs bestimmen kann. Die Wisc-Linien gehörten zu den ersten verfügbaren *Knockout*-Mutanten in *A. thaliana*, was diese aufwändige Strategie erklärt. Später wurden auch für diese Kollektion FSTs generiert [Krysan et al., 1999; Sussman et al., 2000].

Zielsetzung

Ziel dieser Arbeit war ein besseres Verständnis und die Entwicklung von Lösungen für verschiedene, aus der paralogen Natur des *A. thaliana*-Genoms entstehende Probleme sowie die Analyse von Insertionsstellen der T-DNA im Genom von *A. thaliana* auf gemeinsame Merkmale.

Als Basis für diese Arbeiten war es wichtig, dass der Datenbestand der GABI-Kat-Kollektion bezüglich der *A. thaliana*-Genom-Annotation aktualisiert wurde. Damit einhergehend war es von zentraler Bedeutung die genaue Insertionsposition sowie von der Insertion betroffene (und damit in den meisten Fällen defekte) Gene möglichst exakt auf Basis von FST-Sequenzen vorherzusagen. Ein häufig auftretendes Problem in der GABI-Kat-Kollektion waren Kontaminationen der FST-Sequenzen mit DNA aus anderen Linien, sodass eine Insertion nicht eindeutig einer Linie zugeordnet werden konnte. Zur Aufklärung dieser Kontaminationen sollte im Rahmen dieser Arbeit eine Analyse-Strategie entwickelt werden.

Durch paraloge Bereiche innerhalb des *A. thaliana*-Genoms stellte sich in Bezug auf Insertionslinien zunächst die Frage, wie oft die Beobachtung eines Phänotyps bei einem *Knockout* eines Gens durch mindestens eine weitere paraloge Kopie erschwert wird. Insbesondere war von Interesse, wie viele Genpaare es im Genom von *A. thaliana* gibt, bei denen es bereits ausreicht, eine Doppelmutante zu erstellen, um den Phänotyp besser charakterisieren zu können. Die Ergebnisse dieser Analysen sowie auf ihrer Basis erstellte Doppelmutanten sollten abschließend über ein Internetportal der wissenschaftlichen Gemeinschaft zur Verfügung gestellt werden. Ein weiteres Problem mit paralogen Bereichen waren nicht eindeutige Ergebnisse von Sequenzvergleichen bei der FST-basierten Insertionspositions-Vorhersagen. In solchen Bereichen kommt erschwerend hinzu, dass eindeutige Primer zur Aufklärung schwierig zu finden sind. Es sollte eine Methode entwickelt werden, um die Untersuchung von nicht eindeutigen FST-Vorhersagen zu ermöglichen und durch ein geeignetes Primerdesign zu unterstützen.

Auf Basis von *Confirmation*-Sequenzen können gemeinsame Merkmale von T-DNA Insertionen herausgearbeitet werden. Da *Confirmation*-Sequenzen in der Regel den Übergang von der T-DNA in das Kerngenom von *A. thaliana* enthalten, stellen sie dafür ein geeignetes Mittel dar. Insbesondere war eine interessante Fragestellung, ob sich auf Sequenzebene Merkmale finden lassen, die auf die Verwendung von Reparaturmechanismen in *A. thaliana* hindeuten, da eine Beteiligung dieser Mechanismen an der T-DNA Integration vermutet wird.

Ergebnisse

Im Rahmen dieser Arbeit sind vier Publikationen entstanden, die im folgenden Kapitel zusammengefasst und teilweise durch nicht in den Publikationen vorhandene Ergebnisse ergänzt werden. Die Publikationen selbst finden sich im Anhang.

In der ersten Publikation werden verschiedene Verbesserungen des Datenbestandes und der Arbeitsabläufe in der Insertionslinienpopulation GABI-Kat beschrieben. Der größte Schritt war hier die Aktualisierung der FST-Annotation auf TAIRv10 und eine damit einhergehende Optimierung der Insertionspositions-Vorhersage sowie genauere Gen-Annotationen. Eine weitere wichtige Entwicklung war die Etablierung von sogenannten Kontaminationsgruppen. Diese umfassen Vorhersagen, die auf sehr ähnliche Insertionspositionen in verschiedenen Linien hindeuten und dienen dazu, die Linie zu identifizieren in der die Insertion tatsächlich stattgefunden hat.

Nachfolgend liegt der Fokus auf paralogen Bereiche des Genoms von *A. thaliana*. Im GABI-DUPLO-Projekt war das Ziel, eine Kollektion von Linien mit Insertionen in Genpaaren bereitzustellen. Dazu wurde eine Liste von Genen erstellt, die genau ein paraloges Gen im Genom haben. Dafür mussten vorab Auswahlkriterien festgelegt werden. Die erstellten Doppelmutanten wurden auf Phänotypen untersucht und ein Webinterface zur Bereitstellung der Linien für die wissenschaftliche Gemeinschaft implementiert. Die Ergebnisse dieser Arbeiten konnten in einer zweiten Publikation veröffentlicht werden.

Die vermehrte Arbeit mit Insertionen in paralogen Bereichen des Genoms machte es erforderlich, Lösungen für daraus entstehende Probleme zu entwickeln. Paraloge Gruppen wurden eingeführt, um mit nicht eindeutigen Insertionsstellen-Vorhersagen zu arbeiten. Des Weiteren wurde es erforderlich, spezifischere Primer zu generieren. Dazu wurde ein einfach zu benutzendes Werkzeug entwickelt, das die Anzahl der Fehl-Hybridisierungen in paralogen Teilen des Genoms minimiert. Dieses wurde über SimpleSearch einhergehend mit einer dritten Publikation der wissenschaftlichen Gemeinschaft zur Verfügung gestellt.

Die im Rahmen des GABI-Kat Projektes angefallenen *Confirmation*-Sequenzen von bestätigten Linien wurden im Hinblick auf gemeinsame Charakteristika am Übergang von der T-DNA in die genomische DNA untersucht. Dabei zeigte sich, dass viele Gemeinsamkeiten mit nicht-exakten Mechanismen der Reparatur von Doppelstrangbrüchen existieren, und daraus Schlüsse zum Integrationsmechanismus gezogen werden können. Die Ergebnisse dieser Analysen bildeten die Grundlage für eine vierte Publikation.

4.1 Optimierung der Insertionsstellen-Vorhersage und des Umgangs mit Kontaminationen

Publikation 1:

Nils Kleinboelting*, Gunnar Huet*, Andreas Kloetgen, Prisca Viehoveer, and Bernd Weisshaar. GABI-Kat SimpleSearch: New features of the *Arabidopsis thaliana* T-DNA mutant database. *Nucleic Acids Research*, 2012, 40(D1):D1211-D1215

*geteilte Erstautorenschaft

Zusammenfassung:

Entscheidend für eine FST-basierte Kollektion von T-DNA Insertionsmutanten ist eine akkurate Vorhersage von potentiellen Insertionsorten und der Funktion der von der Insertion betroffenen Gene. Dazu ist es notwendig, dass die Annotation der FSTs auf der aktuellsten Annotation des zugrunde liegenden Genoms basiert. Im Rahmen dieser Publikation wurde eine Aktualisierung der FST-Annotationen von TIGR5 auf TAIRv10 vorgenommen. Zusätzlich wurde das Datenmodell der GABI-Kat Datenbank so erweitert, dass auch die Annotierung von mehreren an einem Insertionsort liegenden Genen möglich wurde.

Aufgrund der oftmals schlechten Sequenzqualität von FSTs, insbesondere im 5'-Bereich der Sequenz, ist eine exakte Bestimmung der potentiellen Insertionsposition erschwert. Daher wurde ein optimiertes Verfahren zur Bestimmung der potentiellen Insertionsposition entwickelt. Anstatt den Beginn des besten BLAST-Treffers (Formel 4.1) als potentielle Insertionsposition anzunehmen, wurde zwischen Fällen in denen ein Teil der T-DNA-Sequenz am Beginn der FST-Sequenz gefunden wurde, und Fällen in denen diese nicht gefunden wurde, unterschieden (Formel 4.4). Bei Anwesenheit einer zur T-DNA ähnlichen Sequenz wurde wie bisher verfahren. Bei dessen Abwesenheit wurde anhand der Anzahl der Basen, die im *Tracefile* der zugrunde liegenden Sequenzierung vor Beginn des besten BLAST-Treffers zu finden waren (Formel 4.2) und der Distanz des Sequenzierprimers zur *Border*-Grenze der T-DNA (Formel 4.3) zurückgerechnet. Die Insertionsposition konnte demnach folgendermaßen festgelegt werden (siehe auch *Figure 1* in der oben genannten Publikation):

$$\begin{aligned} \text{BasisInsertionsPosition}(BIP) &= \text{Beginn des besten BLAST-Treffers} & (4.1) \\ & (\text{Subject-Start bei Vorwärts-Treffern, sonst Subject-End}) \end{aligned}$$

$$\text{QueryStartPosition}(QSP) = \text{Query-Start des besten BLAST-Treffers} \quad (4.2)$$

$$\text{Primerabstand}(D) = \begin{cases} 59 & \text{für LB-FSTs mit Primer 8409 sequenziert} \\ 156 & \text{für RB-FSTs mit Primer 3144 sequenziert} \\ 90 & \text{für RB-FSTs mit Primer CR3S sequenziert} \end{cases} \quad (4.3)$$

$$\text{InsPos} = \begin{cases} BIP - QSP + D & \text{wenn T-DNA-Sequenz zu Beginn des FSTs gefunden,} \\ & \text{BLAST-Treffer mind. } D \text{ von Beginn der Sequenz entfernt} \\ & \text{und Vorwärts-Treffer} \\ BIP + QSP - D & \text{wenn T-DNA-Sequenz zu Beginn des FSTs gefunden,} \\ & \text{BLAST-Treffer mind. } D \text{ von Beginn der Sequenz entfernt} \\ & \text{und Rückwärts-Treffer} \\ BIP & \text{sonst} \end{cases} \quad (4.4)$$

Da aufgrund der schlechteren Sequenzqualität im 5'-Bereich des FSTs der Übergang von der T-DNA in die genomische DNA oft nicht gut lesbar ist, wäre der Beginn des besten BLAST-Treffers somit die falsche Insertionsposition. Anhand der Anzahl der zwar unleserlichen aber vorhandenen Basen kann diese Vorhersage wie oben beschrieben verbessert werden. Als Grundlage für diese Berechnungen dienen Sequenzen, die nicht durch Qualitäts-*Trimming* verkürzt wurden.

Einen Vergleich zwischen der ursprünglichen und der wie oben beschriebenen verbesserten Methode, eine Insertionsposition vorherzusagen, zeigt Abbildung 4.1. Der arithmetische Mittelwert der Distanz zur bestätigten Insertionsposition verbesserte sich von 73,7 bp auf 25,8 bp während sich der Median von 23 bp auf 10 bp senken ließ. Wie in der Abbildung zu sehen, ließ sich besonders der Teil der Vorhersagen verringern, der eine Abweichung von 100-300 bp aufweist. Abweichungen in diesem Größenbereich können in Einzelfällen bereits Probleme beim Primerdesign verursachen, wenn mit Primern sehr nah oder sehr weit entfernt von der Insertionsstelle gearbeitet wird.

TAIRv10 bietet neben den Pseudochromosomen anstelle von BAC-Sequenzen als Annotationsbasis auch weitere Fortschritte gegenüber TIGR5. So finden sich auch Informationen über die nicht-translatierten Enden der mRNA (*untranslated region* (UTR)). Des Weiteren sind viele RNA-kodierende Gene und transponierbare Elemente annotiert. Diese Informationen wurden genutzt, um die Definition von Gentreffern in GABI-Kat zu verbessern. Es wurden nun die folgenden Insertionstypen unterschieden:

- **CDSi** - zwischen Start- und Stop-Codon eines Protein-kodierenden Gens oder Pseudogens
- **5'- bzw. 3'-TS2TE** - innerhalb der **annotierten** 5'- bzw. 3'-UTR eines Protein-kodierenden Gens (TS2TE = *transcription start to transcription end*)
- **Promotor-Treffer** - innerhalb von 300 bp vor Beginn der **annotierten** 5'-UTR eines Protein-kodierenden Gens
- **5'- bzw. 3'-Treffer** - innerhalb von 300 bp strangaufwärts bzw. abwärts des Start- bzw. Stop-Codons eines Protein-kodierenden Gens **ohne annotierte** UTRs oder Pseudogens

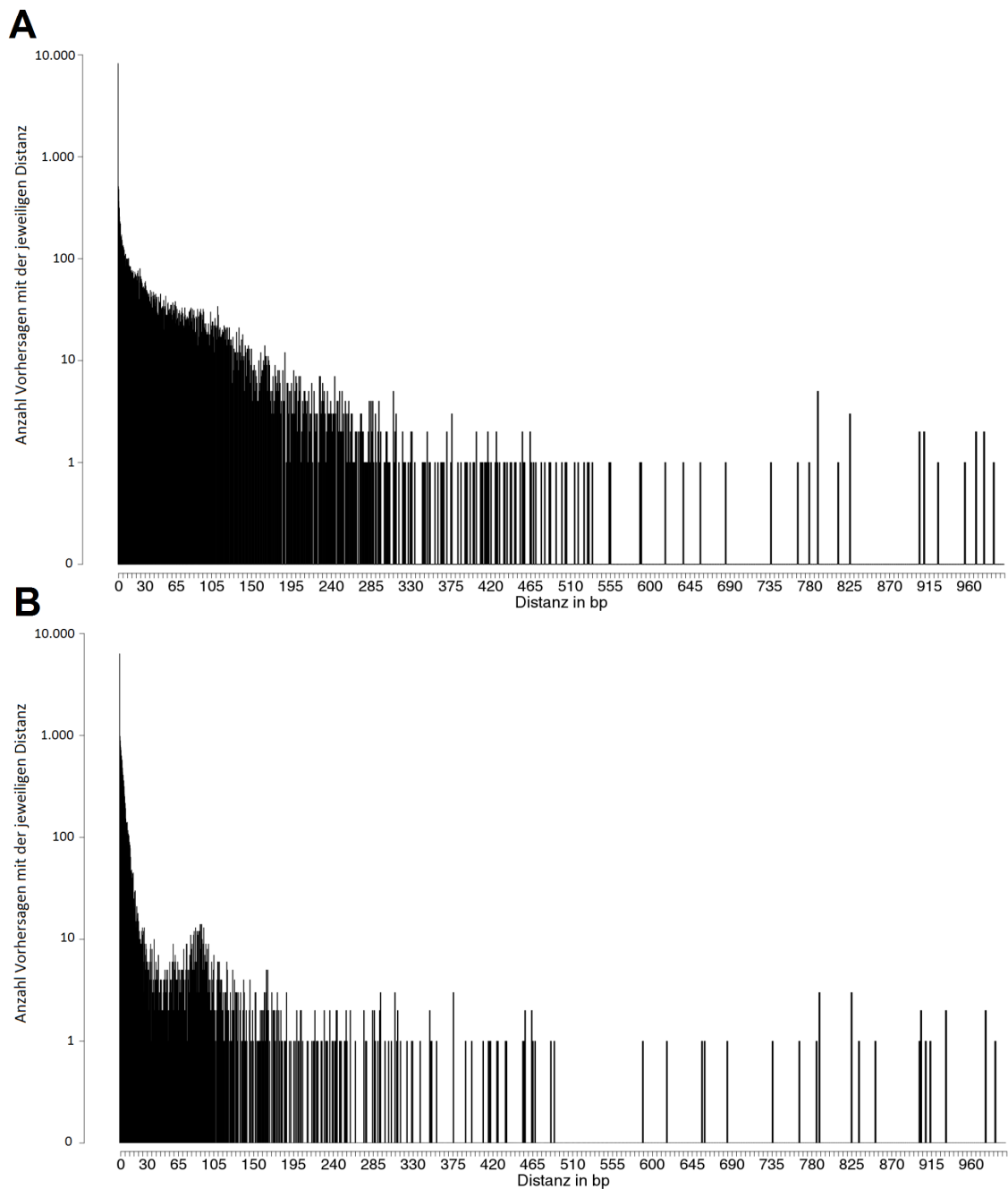


Abbildung 4.1: Abstand zwischen bestätigter und vorhergesagter Insertionsposition nach der alten Methode (A) und der verbesserten Methode (B) in logarithmischer Darstellung.

- **TS2TE** - zwischen Transkriptionsstart und -ende eines RNA-Gens oder transponierbaren Elements

Zusätzlich wurden etwa 20.000 neue FSTs generiert - hauptsächlich für Linien, die bisher noch keine Insertions-Vorhersage besaßen. Damit waren es zum Zeitpunkt

der Publikation etwa 133.000 FSTs in 71.235 Linien und 88.580 Insertions-Vorhersagen. Von den insgesamt 27.416 Protein-kodierenden Genen im Kerngenom von *A. thaliana* sind damit 19.120, d.h. ca. 70 %, allein durch GABI-Kat Insertionen abgedeckt.

Innerhalb der 96er-Blöcke (Einheit in der die T1-Pflanzen angezogen wurden) traten gelegentlich einzelne Blöcke auf, in denen signifikant viele Insertions-Bestätigungen fehlschlagen. Eine weitere Verbesserung der Datenqualität wurde dadurch erreicht, dass diese Blöcke identifiziert und analysiert wurden. Oft lag in diesen Blöcken eine Vertauschung mit Sequenzen aus anderen Blöcken oder eine Drehung des Blocks vor, sodass die Sequenzen der falschen Linie im selben Block zugeordnet waren. Durch die Bereitstellung neuer Analyse-Methoden im GABI-Kat LIMS konnten viele dieser Probleme identifiziert und mit klärenden PCRs aufgelöst werden. Die falschen FST-Zuordnungen wurden korrigiert und die Verlässlichkeit der FST-Vorhersagen insgesamt erhöht.

Um die wertvollsten Insertionen, die sich in der GABI-Kat-Kollektion finden lassen, zu bearbeiten, wurden die Schnittmengen mit anderen großen Kollektionen (SALK, Wisc und SAIL) berechnet. Dadurch konnten diejenigen Insertionen in GABI-Kat identifiziert werden, für die es keine Insertion in den anderen Kollektionen gab. Des Weiteren sind Insertionen in Genen wichtig, für die es in anderen Kollektionen nur genau ein anderes Allel gibt, bzw. kein CDSi-Allel. Diese wertvollen Insertionen sollten daraufhin bevorzugt bestätigt werden mit dem Ziel die entsprechenden Linien an das NASC abzugeben, damit diese auch nach Projektlaufzeit Ende 2014 erhalten bleiben. Insgesamt fanden sich 2.412 einzigartige Insertionen, von denen aufgrund vorangegangener Arbeiten bereits 1.204 bestätigt waren (Ende April 2011). Bis zum jetzigen Zeitpunkt (Ende Februar 2015) konnte die Zahl dieser verifizierten und an das NASC abgegebenen Insertionen auf 1.866 gesteigert werden.

Ein großes Problem bei der Arbeit mit einer Kollektion von Insertionsmutanten sind Kontaminationen der FSTs mit DNA aus anderen Linien. Diese können an verschiedenen Punkten im Arbeitsablauf auftreten. Samen können vertauscht oder vermischt werden. Bei unsauberer Arbeit mit Pipetten kann DNA von einer Zelle der Platte in die Nachbarzelle gelangen. Nicht sorgsam gereinigte Platten führen zu einer Kontamination von einzelnen Zellen in der Sequenzierung der FSTs des darauf folgenden Blockes. All dies führt zu nahezu identischen Insertionspositions-Vorhersagen in mehreren Linien, obwohl alle nur auf genau ein Insertionsereignis zurückgehen.

Zur Lösung dieses Problems wurden sogenannte Kontaminationsgruppen eingeführt. Durch ein hierarchisches *Clustering* werden nahe beieinander liegende Insertionspositionen in verschiedenen Linien zu Gruppen zusammengelegt. Wird eine Insertion innerhalb dieser Gruppen angefragt, werden intern alle Insertionen dieser Gruppe bearbeitet, um zu klären, in welcher Linie sich die Insertion befindet. Wenn möglich, werden für alle PCRs dieselben Primer verwendet. Die so gewonnenen Informationen werden nach erfolgter Bestätigung auch in SimpleSearch zugänglich gemacht. Benutzer werden von nicht bestätigten Insertionen innerhalb einer Kon-

taminationsgruppe zu der bestätigten Insertion weitergeleitet.

Die Einführung von Kontaminationsgruppen brachte ein messbares Resultat: Die Bestätigungsrate, die angibt, wie viele der vorhergesagten Insertionen experimentell verifiziert werden können, wurde um 3 % von 78 % auf 81 % gesteigert. Weitere in dieser Publikation beschriebene Maßnahmen wie eine verbesserte Vorhersage der Insertionsposition und die Korrektur von falsch zugeordneten FSTs brachte eine zusätzliche Steigerung der Bestätigungsrate auf 84 %. Dadurch erhielten Benutzer häufiger die von ihnen bestellte *Knockout*-Mutante, unnötige Laboranalysen wurden vermieden und gezielteres wissenschaftliches Arbeiten ermöglicht.

Die Aktualisierung der FST-Annotation, die Verbesserung der Insertionspositions-Vorhersage sowie die verbesserten Gentreffer-Kriterien bilden die Basis für die nachfolgenden Arbeiten.

4.2 Berechnung von Gruppen paraloger Gene für die Erzeugung von Doppelmutanten in *A. thaliana* und deren Bereitstellung

Publikation 2:

Cordelia Bolle, Gunnar Huep, Nils Kleinboelting, Georg Haberer, Klaus Mayer, Dario Leister and Bernd Weisshaar. GABI-DUPLO: a collection of double mutants to overcome genetic redundancy in *Arabidopsis thaliana*. *The Plant Journal*, 2013(75), 157-171

Zusammenfassung:

Wie in Kapitel 2.1 beschrieben, weist das Genom von *A. thaliana* hauptsächlich aufgrund mehrerer Genomduplikationen viele paraloge Bereiche auf. Deshalb gibt es eine Vielzahl duplizierter Gene mit redundanter oder überlappender Funktion. Im GABI-DUPLO-Projekt war das Ziel daher, *Knockout*-Linien zu erstellen, die Insertionen in einem paralogen Genpaar haben, sogenannte Doppelmutanten. Da ein Phänotyp am wahrscheinlichsten zu beobachten ist, wenn es kein drittes paraloges Gen gibt, das eine ähnliche Funktion erfüllt, war dies eine Voraussetzung für die Generierung von Doppelmutanten.

Um zu berechnen, wie viele und welche solcher Genpaare es im Genom von *A. thaliana* gibt, wurde der in Abbildung 4.2 skizzierte, auf Sequenzähnlichkeit basierte Ansatz, gewählt. Zu unterscheiden sind Genpaare ohne weiteres Homolog, genetisch ungekoppelte Genpaare (mit einem definierten genetischen Mindestabstand) und genetisch ungekoppelte Genpaare, für die eine wahrscheinlich zum *Knockout* führende Insertion (CDSi oder 5'-TS2TE) in GABI-Kat oder SALK verfügbar ist. Letztere sind die sogenannten DUPLO-Genpaare, die im Projekt experimentell bearbeitet wurden.

Als Basis aller Ähnlichkeitsberechnungen dienen die Proteinsequenzen aller Proteinkodierenden Gene in *A. thaliana*. Bei mehreren annotierten *Splice*-Varianten wurde jeweils die Längste verwendet. Um Genpaare zu finden, wurde zunächst ein Ähnlichkeitsgraph erstellt. In diesem Graphen stellt jedes Protein einen Knoten dar. Eine Kante zwischen zwei Knoten existiert, wenn die korrespondierenden Proteine ausreichend ähnlich zueinander sind. Als ausreichend ähnlich wurde eine Ähnlichkeit von mindestens 60 % der Aminosäuren und eine maximale Lückengröße von 20 % des optimalen globalen *Alignments* beider Proteinsequenzen festgelegt. Der Needleman-Wunsch-Algorithmus ist geeignet, um ein optimales globales *Alignments* zweier Sequenzen zu berechnen. Allerdings ist er, im Gegensatz beispielsweise zur Heuristik BLAST für lokale *Alignments*, relativ rechenaufwändig und die Anzahl der zu berechnenden *Alignments* für alle Kombinationsmöglichkeiten von mehr als 27.000 Proteinen in *A. thaliana* sehr hoch. Um die Anzahl der zu berechnenden *Alignments* zu reduzieren, wurden alle diejenigen Kombinationen ausgeschlos-

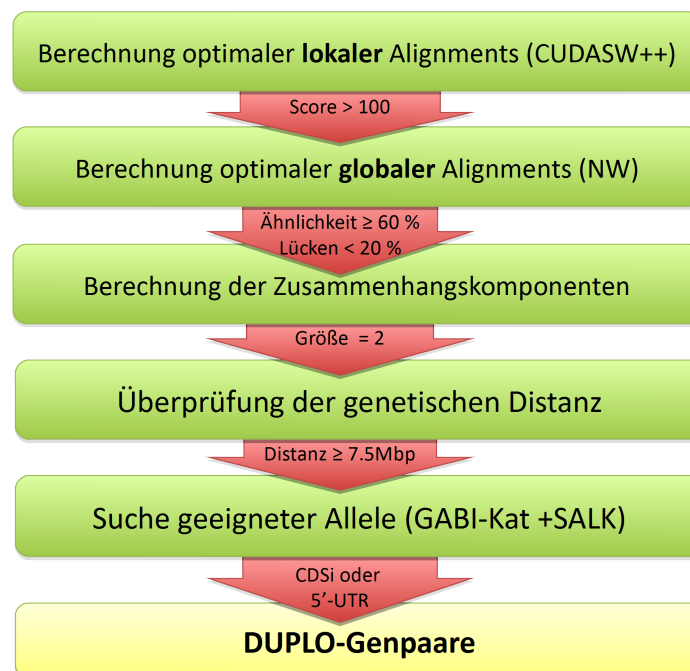


Abbildung 4.2: Workflow zur Berechnung der DUPLO-Genpaare. Globale optimale *Alignments* der Proteinsequenz von vorher durch lokale *Alignments* gefilterten Kandidaten dienen als Basis für einen Graphen, in dem Knoten (Gene) verbunden sind, wenn die Ähnlichkeit der von ihnen kodierten Proteine ausreichend hoch ist. Die Zusammenhangskomponenten, die genau aus zwei Genen bestehen, sind Genpaare. Um aussichtsreiche Kandidaten für eine Kreuzung zu erhalten, ist eine minimale genetische Distanz der beteiligten Gene, sowie die Verfügbarkeit eines *Knockout*-Allels erforderlich.

sen, die keine ausreichende Ähnlichkeit in einem lokalen *Alignment* haben. Diese lassen sich verhältnismäßig schnell berechnen. Für optimale lokale *Alignments* eignet sich der Smith-Waterman-Algorithmus, für den auch mit CUDASW++ eine CUDA-Implementierung verfügbar ist [Liu et al., 2009]. Durch Verteilung der Berechnungen auf mehrere Tesla-Grafikprozessoren war es möglich, die nötigen Berechnungen innerhalb eines Tages durchzuführen. Für alle Proteinpaare mit einem *Score* von mindestens 100 wurden daraufhin optimale globale *Alignments* mittels Needleman-Wunsch (NW) und der Distanzmatrix BLOSUM62 [Henikoff and Henikoff, 1992] berechnet.

Anhand der berechneten Werte zu Protein-Ähnlichkeiten konnte der Ähnlichkeitsgraph wie oben beschrieben erstellt werden. Um Genpaare zu identifizieren, wurden anhand dieses Graphen alle Zusammenhangskomponenten berechnet. Eine Zusammenhangskomponente ist ein Teilgraph des Gesamtgraphen, in dem jedes Paar von

Knoten über mindestens einen Kantenpfad miteinander verbunden ist. Gesucht ist dabei der maximale Teilgraph, d.h. es gibt keine weiteren Knoten, die mit diesem Graphen verbunden sind (siehe Abbildung 4.3). Wichtig ist der Unterschied zu einem vollständigen Teilgraphen, in dem alle Knoten durch Kanten verbunden sind. So reicht es bei Verwendung der Zusammenhangskomponenten-Definition aus, dass es ein drittes Homolog gibt, das nur zu einem von beiden Proteinen die erforderliche Sequenzähnlichkeit aufweist, um eine Zusammenhangskomponente der Größe drei zu bilden und das Tripel aus der Menge von Genpaaren auszuschließen. Für die Berechnung wurde ein einfacher Tiefensuche-Algorithmus implementiert, der in linearer Zeit alle Zusammenhangskomponenten eines Graphen berechnet.

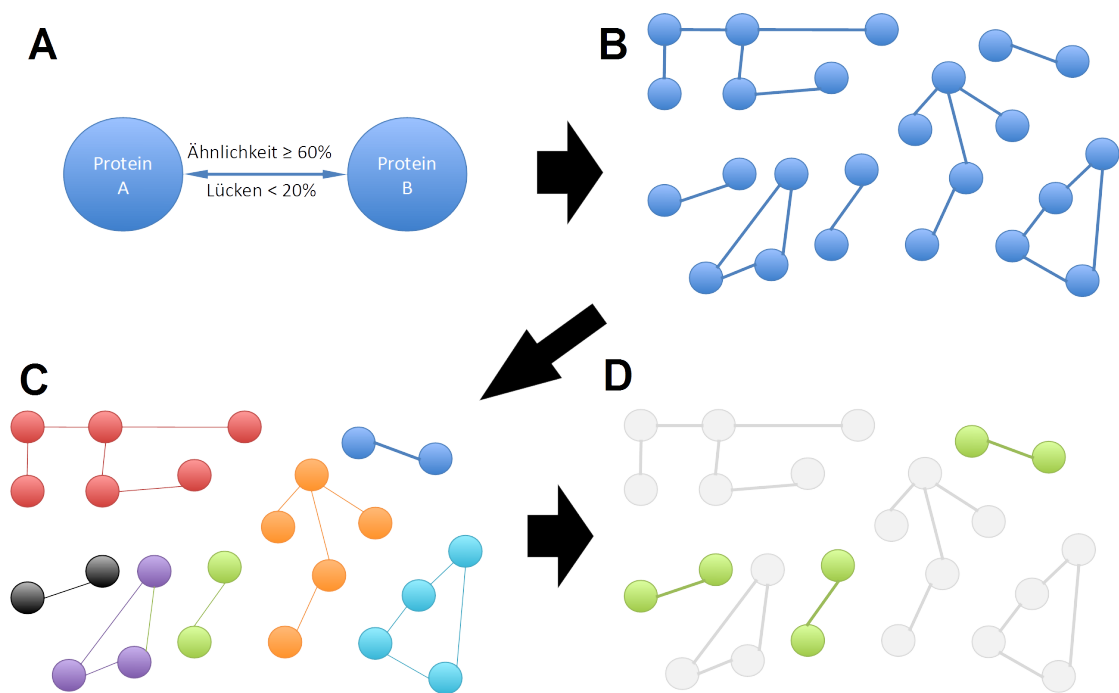


Abbildung 4.3: Berechnung von Zusammenhangskomponenten der Größe zwei. Auf der Basis von globalen optimalen *Alignments* von Proteinpaaren mit ausreichend Ähnlichkeit zueinander (A) wird ein Ähnlichkeitsgraph erstellt (B). Die Zusammenhangskomponenten werden berechnet, wobei eine Zusammenhangskomponente dann abgeschlossen ist, wenn keine weiteren Verbindungen zu anderen Knoten mehr existieren (C). Diejenigen Zusammenhangskomponenten mit Größe zwei sind Genpaare ohne ein weiteres Homolog (D) und bilden die Basis für DUPLO-Genpaare.

Das Ergebnis dieser Berechnungen ist in Tabelle 4.1 dargestellt. Der überwiegende Anteil der berechneten Proteinfamilien besteht aus zwei bis fünf Proteinen. Etwas über die Hälfte aller Proteine ist nicht Teil einer entsprechend berechneten Proteinfamilie.

Tabelle 4.1: Verteilung der Zusammenhangskomponenten.

Größe der Zusammenhangskomponente	Anzahl Zusammenhangskomponenten
1	13676
2	2594
3	697
4	312
5	161
6	98
7	66
8	57
9	43
10	32
11	25
12	15
13	19
14	13
15	9
16	2
17	7
18	8
19	6
20 und größer bis zu 50	31

Anschließend werden die gefundenen Genpaare weiter auf ihre genetische Distanz gefiltert. Zur Erzeugung von Doppelmutanten müssen die beiden Eltern-Linien gekreuzt werden. Damit dies möglich ist, müssen beide Insertionen, falls sie auf demselben Chromosom liegen, eine gewisse genetische Distanz aufweisen, damit ein Crossing-Over Ereignis wahrscheinlich ist. Ein Centimorgan (cM) ist die genetische Distanz die einer 1 %-igen Chance entspricht, dass zwei Stellen dieser Distanz im Genom innerhalb einer Generation durch Rekombination während der Meiose getrennt werden. Die Länge eines cM in *A. thaliana* ist 217 kbp [Mezard, 2006]. Die in GABI-DUPLO gewählte minimale Distanz ist 7,5 Mbp was einer etwa 35 %-igen Chance eines für eine Doppelmutante erforderlichen Crossing-Overs entspricht. Die Zentromere wurden dabei behandelt wie 1000 bp, da eine Rekombination in diesem Bereich unwahrscheinlicher ist.

In einem letzten Schritt wurde in der GABI-Kat und in der SALK-Kollektion nach Insertionen in den beiden Partner-Genen gesucht. Dabei wurden bevorzugt CDSi-Insertionen ausgewählt. Falls in beiden Kollektionen keines vorhanden war, wurde auf 5'-TS2TE Insertionen ausgewichen.

Von allen 370.069.615 möglichen Kombinationen für Genpaare in *A. thaliana* weisen 26.982 Paare die erforderliche Homologie auf (Ähnlichkeit $\geq 60\%$, Lücken $< 20\%$). Nach Filterung auf Zusammenhangskomponenten der Größe zwei verblieben

noch 2.594 Genpaare. Davon waren 2.108 genetisch ungekoppelte Paare, von denen sich für 1.294 ein CDSi- oder 5'-TS2TE-Allel finden ließ.

Die Eltern-Linien wurden in Bielefeld bestätigt, genotypisiert und wenn möglich in Form von homozygoten Samen zum Projektpartner nach München geschickt, dessen Aufgabe es war, die Kreuzungs-Linien zu erstellen. Nach erfolgter Kreuzung wurden in der F2-Generation Nachkommen identifiziert, die homolog für beide Insertionsallele sind. Die anfallenden Daten wurden im GABI-Kat LIMS gespeichert und verwaltet, wozu das Datenschema und die Benutzeroberfläche entsprechend erweitert wurden. Zum Ende der Projektlaufzeit im Dezember 2011 waren 234 Linien für DUPLO-Genpaare erstellt und beim NASC verfügbar, in die Publikation wurden davon 200 aufgenommen. Für einen Großteil der übrigen Paare wurden die jeweiligen Einzel-Insertionen bestätigt, sodass interessierte Forscher die für die Kreuzung benötigten Linien beim NASC bestellen können.

Zur Bereitstellung der gewonnenen Informationen über Genpaare im Genom von *A. thaliana* sowie Bearbeitungsdetails zu einzelnen Linien oder Doppelmutanten wurde im Rahmen dieser Arbeit das Webinterface DUPLOdb (*DUPLO database*) implementiert, das Zugriff auf und Suche in diesen Daten ermöglicht [DUPLOdb, 2015]. Abbildung 4.4 zeigt exemplarisch die Darstellung eines einzelnen DUPLO-Genpaares über die Webseite.

Von den in der Publikation untersuchten 200 Doppelmutanten zeigten 13 einen direkt sichtbaren, vom Wildtyp abweichenden Phänotyp. In 23 Linien schien der *Knockout* eine Letalität der Doppelmutante zu bewirken und 14 Doppelmutanten sind für einen Phänotyp unter bestimmten Bedingungen aus der Literatur bekannt. Einen beobachtbaren Phänotyp zeigen also 50 Linien (25 %), was die Nützlichkeit von Doppelmutanten für die Zuweisung von Genfunktionen unterstreicht.





GABI DUPLodb - pair details			
Details for pair 932			
	Pair Info	Parent 1	Parent 2
Pair ID	932		
Parent loci		At3g57420 [TAIR] [MIPS] [SIGNAL]	At2g41770 [TAIR] [MIPS] [SIGNAL]
Gene Annotations		Protein of unknown function (DUF288)	Protein of unknown function (DUF288);
Pair-related paper(s)	[Bolte et al. 2013]		
Similarity (TAIRv10)	93.7 %		
Gap (TAIRv10)	1 %		
Component	2		
Genetically unlinked	Yes		
Assigned allele		100H05 / At3g57420	733B10 / At2g41770
Insertion classification		CDSi	CDSi
Parent line's graphic view			
Parent line's confirm. status		confirmed, show confirmation seq's	confirmed, show confirmation seq's
Primer and wt-amplicons		show primer details	show primer details
Doublemutant ID	932/1b2.38.1.1		
NASC set ID	N2102617		
Assigned allele		100H05	733B10
Allele genotype in parent1		Hom	
Allele genotype in parent2			Hom
Allele genotype in DM1		Hom	Hom
Observed phenotype	dwarfish		
Image(s) (click image to enlarge)			
	932 DM with parental lines		932 DM detail

Abbildung 4.4: Das DUPLO-Genpaar Nummer 932 in DUPLodb. Details zu beiden Eltern-Linien, Links zu Primern und *Confirmation*-Sequenzen, sowie Informationen zum Genotyp der Kreuzungslinie werden dargestellt. Über die NASC-ID kann die Linie direkt beim NASC bestellt werden. An den Fotos lässt sich erkennen, dass in diesem Fall nur ein Phänotyp zu beobachten ist, wenn beide Gene ausgeschaltet sind (932DM). Screenshot von DUPLodb [2015].

4.3 Entwicklung eines Primerdesigns optimiert für paraloge Bereiche des Genoms von *A. thaliana*

Publikation 3:

Gunnar Huep*, Nils Kleinboelting* and Bernd Weisshaar. An easy-to-use primer design tool to address paralogous loci and T-DNA insertion sites in the genome of *Arabidopsis thaliana*. *Plant Methods*, 2014, 10:28

*geteilte Erstautorenschaft

Zusammenfassung:

Im GABI-DUPLO-Projekt wurden Insertionen in paralogen Bereichen des Genoms bearbeitet. Das Hauptproblem, das dadurch entstand, war sicherzustellen, dass jeweils das korrekte paraloge Gen untersucht und bestätigt wurde. Deshalb war es zunächst wichtig, die FST-basierte Insertionsstellen-Vorhersage zu erweitern, sodass zu einem FST auch mehrere Treffer abgeleitet wurden. Des Weiteren mussten die auf diese Weise vorhergesagten Insertionen mittels einer klärenden PCR untersucht werden, bei der Primer zum Einsatz kommen, die möglichst nur ein einziges Amplimer im Genom bilden.

Nach einer Erweiterung des Datenbank-Modells und der beiden Benutzerschnittstellen war es möglich, mehr als eine BLAST-Vorhersage für jeden FST zu speichern. Zu unterscheiden sind hier zum Einen Vorhersagen aufgrund von zusammengesetzten FSTs und zum Anderen Vorhersagen aufgrund von paralogen Bereichen. Durch die zufällige Position der genomischen *BfaI*-Schnittstelle während der FST-Generierung ist die Größe des Amplikons, das zur Sequenzierung kommt, unterschiedlich. Gibt es mehr als eine Insertion in einer Linie, gibt es auch mehrere Fragmente mit unterschiedlicher Größe, die gemeinsam sequenziert werden. Dadurch, dass von einem kürzeren Fragment tendenziell mehr Produkt gebildet wird, liefert dessen Sequenz zu Beginn des FSTs ein stärkeres Signal, gefolgt vom schwächeren Signal der Sequenz des längeren Fragments (siehe Abbildung 4.5). Ein einzelner FST kann somit Hinweise auf mehrere Insertionen geben. Sie sind dadurch gekennzeichnet, dass unterschiedliche Bereiche desselben FSTs BLAST-Vorhersagen an unterschiedlichen Stellen des Genoms ergeben. Diese Teile werden im Folgenden als „Regionen“ bezeichnet. Vorhersagen aufgrund von paralogen Bereichen ergeben sich aus ähnlich guten BLAST-Ergebnissen für dieselbe Region eines FSTs.

Zur Klassifizierung der einzelnen Vorhersagen wurden diese in Kategorien aufgeteilt: Treffer der Kategorie 0 entsprechen dabei der besten Vorhersage und sind in SimpleSearch zugänglich. Kategorie 1 beinhaltet Treffer mit einem guten *e-Value*, einer nicht zu großen Differenz zum besten Treffer, sowie weitere Treffer aufgrund von mehreren Regionen (siehe Methoden-Teil in der Publikation für Details). Kategorie 2 beinhaltet weitere Treffer mit schlechtem *e-Value*.

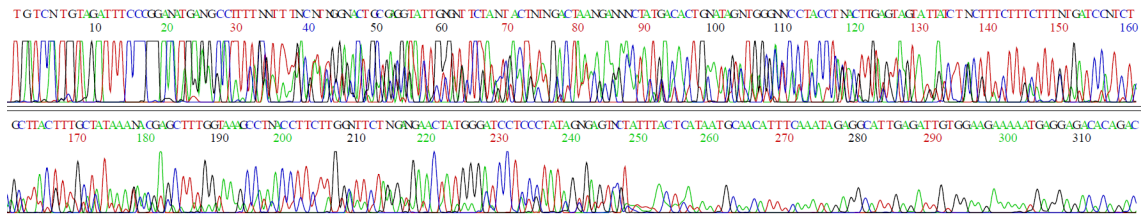


Abbildung 4.5: Tracefile des FSTs 28-K012311-022-110-D04-8409 in der Linie 110D04 sequenziert mit dem *Border*-spezifischen Primer 8409. Im Bereich bis ca. 120 bp sind zwei starke sich überlagernde Signale zu beobachten. Darauf ist nur das Signal des längeren Fragments zu erkennen. Dieser FST deutet auf zwei Insertionen hin. Der Bereich von 1-83 bp ist einer Insertion auf Chr2:2146664 in *At2g05715* zuzuordnen, der Bereich von 121-190 bp einer Insertion auf Chr3:20416351 in *At3g55090*. Beide Insertionen konnten experimentell bestätigt werden. Über die Gesamtlänge des FSTs findet sich ein schwaches drittes Signal, das den Qualitätsanforderungen bei der Prozessierung nicht genügt.

Treffer der Kategorien 0 und 1 wurden zu paralogen Gruppen zusammengefasst, wenn sie von derselben Region stammen und ähnliche Bereiche im Genom von *A. thaliana* treffen [Kloetgen, 2011]. Nach der Bestellung eines Benutzers konnten so alle Insertionen innerhalb einer paralogen Gruppe untersucht werden.

Durch die Erweiterung auf mehrere BLAST-Treffer pro FST gab es einen Zuwachs an vorhergesagten Insertionen. Mit nur je einem gespeicherten Treffer gab es bei 135.210 FSTs mit BLAST-Treffer gegen die *A. thaliana*-Genomsequenz insgesamt 91.383 vorhergesagte Insertionen. Aus diesen FSTs abgeleitet gab es 153.919 Regionen, deren beste Treffer 102.494 Insertionen voraussagen. Von diesen Regionen wiederum haben 137.985 genau einen Treffer im *A. thaliana*-Genom, für 15.934 wurden zwei oder mehr Treffer gespeichert. Mehrere Vorhersagen für dieselbe Insertion zusammengenommen ergeben 94.260 Insertionen bei den eindeutigen Regionen bzw. 38.038 Insertionen unter den Regionen mit paralogen Vorhersagen. Es wurden 10.737 paraloge Gruppen gebildet, in denen 28.836 Insertionsstellen-Vorhersagen enthalten sind. Von diesen enthalten 4.605 Gruppen Insertionen mit sehr ähnlichen Treffern. In diesen Gruppen ist es normalerweise nötig, mehrere Insertionen, möglichst mit einzigartigen Primern zu untersuchen. In 1.139 von 1.609 aufgelösten Gruppen konnte eine Insertion bestätigt werden. Die Differenz stellt Gruppen dar, die auf Kontaminationen zurückzuführen sind. In der überwiegenden Anzahl der Fälle (964 von 1.139) war die bestätigte Insertion identisch mit der besten Vorhersage. In 14 % (160) der paralogen Gruppen, in denen eine Insertion bestätigt werden konnte, gab es mindestens eine zweite gleich gute Vorhersage, von denen nur eine korrekt war. Zusätzlich gab es eine kleine Anzahl (15) paraloger Gruppen, in denen schlechtere Vorhersagen bestätigt wurden.

Um bei der Untersuchung von Insertionen mittels PCR die möglichen Produkte zu minimieren wurde ein Werkzeug zur Generierung optimaler Primer entwickelt.

Zunächst war es wichtig, die Anzahl der Fehlhybridisierungen zu minimieren. Da in paralogen Bereichen des Genoms oft eine Vielzahl von Primern überprüft werden muss, musste das Verfahren dazu schnell berechenbar sein, damit Primer in kurzer Zeit generiert werden können.

In der Laborpraxis hatte sich herausgestellt, dass bereits kurze Sequenzähnlichkeit am 3'-Ende des Primers genügt, um ein Produkt zu erzeugen. Aus diesem Grund ist es nicht ausreichend, wenn die komplette Primersequenz im Genom einzigartig ist. Um Fehlpaarungen zu verhindern, wurde versucht, die Hybridisierungsmöglichkeiten des 3'-Endes zu minimieren. Um dies schnell und effektiv für einen gegebenen Primer zu überprüfen wurde ein Index berechnet, in dem für jedes 12-mer (eine Sequenz der Länge 12) gespeichert wurde, wie oft es im Genom vorkommt. Die Länge von 12 bp wurde deshalb gewählt, weil es zum Einen die kleinste beobachtete Paarung in der praktischen Arbeit mit *A. thaliana* war und weil sie zum Anderen ein guter Kompromiss war zwischen benötigtem Speicher und der Möglichkeit, überhaupt einen Primer mit einzigartigem 3'-Ende zu finden. Abbildung 4.6 zeigt, wie viele 12-mer eine bestimmte Anzahl an Treffern im Genom aufweisen.

Insgesamt wurden 15.626.587 12-mer untersucht. Davon sind lediglich 1.413.976 einzigartig, also etwa 9 %. Des Weiteren haben 8.032.018, also etwa 51,4 %, zwischen 2 und 10 Treffer. Dies zeigt, dass ein zufällig gewählter Primer im Genom in den meisten Fällen kein einzigartiges 3'-Ende hat. Meistens ist es jedoch trotzdem

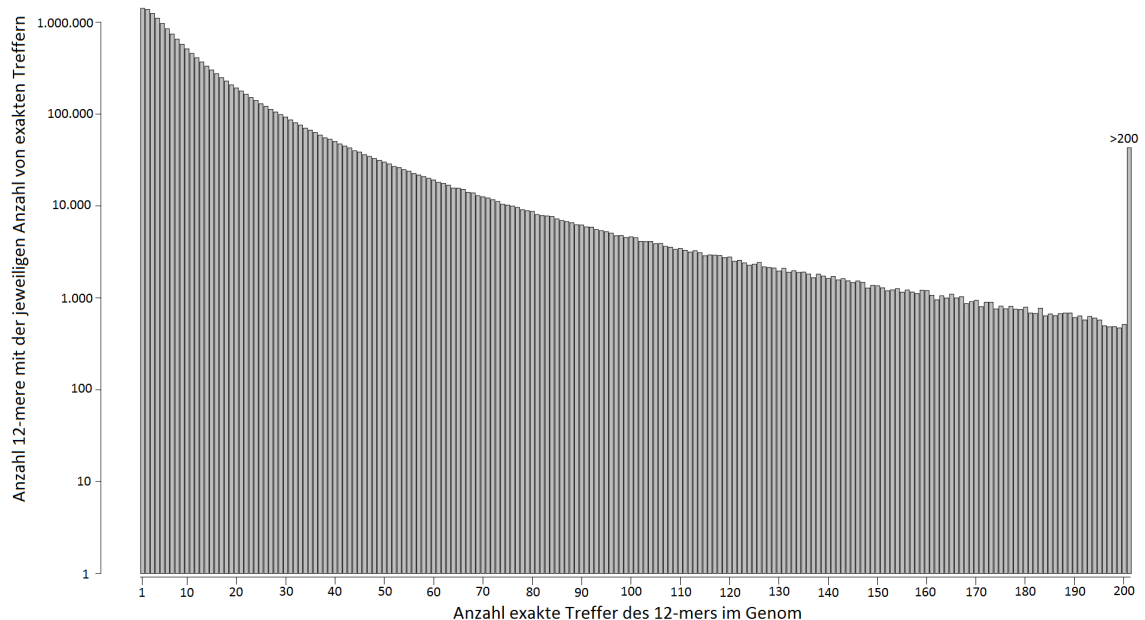


Abbildung 4.6: Häufigkeits-Verteilung aller 12-mer im Genom von *A. thaliana*. Sequenzen, die genau einmal im Genom vorkommen, sind zwar am häufigsten, allerdings haben etwa 91 % der 12-mer 2 oder mehr Treffer.

möglich, einen einzigartigen Primer für eine Zielregion zu finden. Falls nicht, ist das Ziel, die Anzahl der theoretischen Hybridisierungsmöglichkeiten zu minimieren.

Zur Generierung geeigneter Kandidaten für eine gegebene Position im Genom von *A. thaliana* wurden zwei Methoden implementiert. Je nach Beschaffenheit der Zielregion, in der der Primer liegen soll, kommen zwei verschiedene Strategien zum Einsatz: Die Primer3-basierte Methode und die paraloge Primer Methode.

In der Primer3-basierten Methode werden mittels Primer3 [Untergasser et al., 2012] iterativ verschiedene Primer generiert, bis derjenige mit den wenigsten Hybridisierungsmöglichkeiten identifiziert wurde. In Bereichen, zu denen sich viele homologe Bereiche im Genom von *A. thaliana* finden, kommt die paraloge Primer Methode zum Einsatz, die über ein multiples *Alignment* und eine Suche nach einzigartigen Sequenzen anhand dieses *Alignments* funktioniert.

Um zu entscheiden, welche von beiden Methoden im jeweiligen Fall Anwendung findet, wird zunächst eine BLAST-Suche der Zielregion (200-1450 bp Entfernung zur Insertion) gegen das komplette Genom von *A. thaliana* mit einem *e-Value*-Schwellenwert von $1e^{-5}$ durchgeführt. Gibt es nach Auswertung aller Treffer eine Teilsequenz der Zielregion, die keinen anderen BLAST-Treffer im Genom hat und mindestens 100 bp lang ist, wird die Primer3-basierte Methode verwendet, ansonsten die paraloge Primer Methode.

Primer3-basierte Methode Der zuvor identifizierte einzigartige Bereich innerhalb der Zielregion wird verwendet, um verschiedene Primer mit dem Programm Primer3 zu generieren. Dazu wird iterativ ein Fenster über den Zielbereich geschoben (das jeweils um eine Primerlänge mit dem vorherigen überlappt) und in jedem davon einen Primer generiert [Zekic, 2012]. Dies wird mit leicht modifizierten Zielwerten für die *Annealing*-Temperatur wiederholt, bis ein Primer gefunden wurde, der ein einzigartiges 3'-Ende hat. Sind alle möglichen Fenster und Temperatur-Zielwerte getestet und kein solcher Primer gefunden, wird dieselbe Prozedur noch einmal mit der kompletten Zielregion wiederholt. Dabei werden sehr kleine und sehr große Distanzen zur Insertionsposition vermieden. Wenn am Ende dieser Prozedur ebenfalls kein einzigartiger Primer gefunden wurde, wird derjenige als Resultat geliefert, der die wenigsten Treffer im Genom von *A. thaliana* ergab.

Paraloge Primer Methode Bei dieser Methode wird die komplette Zielregion zunächst in überlappende Teilregionen aufgeteilt. Zu jeder davon werden anhand einer BLAST-Suche paraloge Bereiche identifiziert, die dann auf die Länge der Teilregion verlängert und mittels ClustalW [Thompson et al., 2002] zu einem multiplen *Alignment* zusammengefasst werden. Innerhalb jedes dieser multiplen *Alignments* wird nun nach Positionen gesucht, die möglichst viele Fehlpaarungen zu der Zielregion aufweisen. Die Fehlpaarungen werden als 3'-Ende verwendet und verlängert, bis die gewünschte *Annealing*-Temperatur des Primers erreicht ist [Kloetgen, 2011]. Schließlich muss der Primer noch diverse Qualitätskriterien wie die Vermeidung von

Sekundärstrukturen oder längerer Abschnitte derselben Base erfüllen (siehe Publikation für Details). Alle erhaltenen Primer werden wiederum auf Einzigartigkeit überprüft und der bestmögliche Primer als Ergebnis geliefert.

Das entwickelte Programm wurde intern für das Primerdesign in GABI-Kat verwendet und fortwährend optimiert. Mit der Publikation wurde es auch durch Implementierung einer sehr einfach zu benutzenden Webseite für die wissenschaftliche Gemeinschaft zugänglich gemacht (siehe Abbildung 4.7). Es erfordert nur eine minimale Parameterauswahl des Benutzers, wie die genaue Position, auf die die Primer zielen sollen, minimaler und maximaler Abstand zu dieser Position sowie eine Ziel-Temperatur. Sollten die beim ersten Versuch generierten Primer nicht für den Benutzer akzeptabel sein, können weitere alternative Primer generiert werden.

2 T-DNA insertion predictions/alleles that hit this region of BAC F21E10 are displayed.

Chr5:9343837 Chr5:9346662 Chr5:9350287

At5g26650 At5g26640

1000 base pairs

Zoom in Zoom out Insertion view

Jump to position where primers should be designed: Chromosome 5 Position: 9346662 Update

Jump to gene code: Update

▼ Confirmed	► Genes (encoding proteins)	▼ Outdated insertion prediction
▼ Not studied, can be ordered	► Genes (for ncRNA/miRNA)	► Confirmation sequence(s)
▼ Failed	► Genes (transposable elements)	► FST(s)
		► Primer

Primer design

TM: 60.5 MIN distance: 300 MAX distance: 800

Design amplicon (primer pair) for selected genome position Design an additional primer pair

Primers shown in visualisation:

forward primer: ATTATGTAA AAAATCGAGGGTA with length 23 nt and TM 62.85 °C; position on Chr5: 9346196 ==> 9346218
reverse primer: CTTGGCTAATCTCTTAGGGTGAG with length 24 nt and TM 59.82 °C; position on Chr5: 9347126 <== 9347149
Size of this amplicon: 954 bp. The amplicon spans a GK insertion position, check [HERE](#) for genotyping amplicon suggestions.

Abbildung 4.7: Das Primerdesign-Tool in SimpleSearch. In diesem Beispiel wurde ein Primerpaar auf Chromosom 5 für eine GABI-Kat Insertion generiert. Das Primerpaar eignet sich sowohl zur Bestätigung der Insertion in Kombination mit *Border*-spezifischen Primern als auch zur Genotypisierung. Über einen Tooltip („HERE“) werden die berechneten Informationen zu Amplicon-Größen angezeigt. Screenshot von der GABI-Kat Webseite [2015].

4.4 Bioinformatische Analyse von *Confirmation*-Sequenzen in Bezug auf Merkmale des Integrationsmechanismus

Publikation 4:

Nils Kleinboelting*, Gunnar Huep*, Ingo Appelhagen, Prisca Viehovever, Yong Li and Bernd Weisshaar. Evaluation of the structural features of thousands of T-DNA insertion sites indicates a double-strand break repair based insertion mechanism. *Molecular Plant*, submitted

*geteilte Erstautorenschaft

Zusammenfassung:

Der Integrationsmechanismus von *Agrobacterium tumefaciens* ist bis zur Integration der T-DNA in das Genom von *A. thaliana* verhältnismäßig gut untersucht. Den verbreitetsten Modellen nach integriert die T-DNA als doppelsträngiges Molekül in einen in der Wirts-DNA vorhandenen Doppelstrangbruch (siehe Kapitel 2.2.3). Um zu untersuchen, ob und welche Reparaturmechanismen dabei zum Einsatz kommen, kann die Schnittstelle zwischen T-DNA und genomischer DNA genauer untersucht werden. Reparierte Doppelstrangbrüche weisen je nach verwendetem Mechanismus einige charakteristische Merkmale wie Deletionen, Mikrohomologien zwischen den beiden reparierten Enden oder *Filler*-Sequenzen auf.

Ein Datensatz für solche Analysen stellt die Sammlung von *Confirmation*-Sequenzen in GABI-Kat dar. Diese Sequenzen zeichnet aus, dass sie in den meisten Fällen den Übergang zwischen T-DNA und genomischer DNA abdecken und gleichzeitig eine hohe Sequenzqualität im Vergleich zu oft qualitativ schlechten FST-Sequenzen aufweisen. Während der Projektlaufzeit von GABI-Kat wurden über 34.000 *Confirmation*-Sequenzen für GABI-Kat-Linien generiert. Zusätzlich dazu gibt es etwa 2.300 Sequenzen, die aus der Untersuchung von SALK-Linien während der Erstellung der DUPLO-Linien stammen. Diese Sequenzen bildeten den Ausgangspunkt für eine Reihe von Analysen spezieller, Sequenz-basierter Charakteristika von T-DNA Insertionen.

Die genauen Schnittstellen der innerhalb von *A. tumefaciens* durch Vir-Proteine herausgeschnittenen und in der Pflanzenzelle ggf. weiter verkürzten T-DNA innerhalb des verwendeten Vektors wurden untersucht. Ein bei der Reparatur von Doppelstrangbrüchen oft beobachtetes Merkmal sind Mikrohomologien und *Filler*. Von Mikrohomologie spricht man, wenn im Übergang von T-DNA zur genomischen DNA ein kurzer Sequenzteil zu beiden Ursprüngen passt. Überlappen die beiden Sequenzbereiche dagegen nicht und es findet sich ein kurzer Sequenzabschnitt dazwischen, der zu keinem von beiden passt, ist ein sogenannter *Filler* vorhanden. Diese *Filler* wurden weiter auf ihren Ursprung und auf Mikrohomologie zur benachbarten Sequenz (T-DNA oder *A. thaliana*-Kerngenom) untersucht. Eine Reihe von Insertionen wurde von beiden Flanken untersucht, um Veränderungen der Ur-

sprungssequenz an der Insertionsstelle zu ermitteln. Zur Analyse der *Confirmation*-Sequenzen wurde die in Abbildung 4.8 skizzierte *Pipeline* implementiert. Ausgehend von *Confirmation*-Sequenzen liefert sie Ergebnisse zu den oben beschriebenen Charakteristika einer T-DNA Insertion.

4.4.1 Untersuchung von *Border*-Schnittstellen

Für die Analysen von *Border*-Schnittstellen sowie nachfolgende Auswertungen wurden insgesamt 30.507 *Confirmation*-Sequenzen berücksichtigt. Nach diversen Schritten zum Ausschluss nicht geeigneter Sequenzen (siehe Abbildung 4.8 oder Publikation) gingen 11.802 Sequenzen in die Analysen ein, wobei jede davon eine Flanke einer Insertion repräsentiert. Von der *Right Border* (RB) stammen dabei 699 Sequenzen, die übrigen 11.103 von der *Left Border* (LB). Die erwartete Schnittstelle der T-DNA liegt bei -3 bp für die LB und bei -22 bp für die RB. Diese beiden Schnittstellen stellen auch das Maxima in den Verteilungen für beide *Border*-Datensätze dar. Meistens ist die T-DNA aber an beiden Seiten weiter verkürzt (72,3 % für LB, 68,2 % für RB). In wenigen Fällen wurde auch ein Teil des Vektors integriert. Das entspricht 5,6 % der Fälle für LB und 8,3 % für RB mit einem Median von 13 bzw. 21,5 bp. Der Median der Differenz zu der erwarteten Schnittposition für LB liegt bei -7 und bei -2 für RB.

Die Ergebnisse zeigen, dass die meisten Schnitte an der Erkennungssequenz liegen oder die T-DNA weiter verkürzt wurde. Beide *Border* scheinen prinzipiell gleich prozessiert zu werden, tendenziell wird die LB dabei etwas weiter verkürzt als die RB.

4.4.2 Untersuchung von Mikrohomologien und *Fillern*

Die Analyse der 11.802 *Confirmation*-Sequenzen ergab, dass die meisten von ihnen Mikrohomologie oder *Filler* aufweisen. Mikrohomologien konnten in 45,4 % der untersuchten Fälle beobachtet werden, *Filler* dagegen in 48,7 %. Der größte Teil der Mikrohomologien lag mit 84,6 % bei bis zu 5 bp, die verbleibenden Fälle wiesen Mikrohomologie von in Einzelfällen bis zu 20 bp auf. Der Median lag bei 3 bp. *Filler* waren meistens zwischen 1 und 10 bp groß (61,6 % der untersuchten Sequenzen), zwischen 10 und 20 bp fanden sich 20,1 % und 18,3 % waren 20 bp oder größer. *Border*-spezifische Unterschiede bezüglich der Mikrohomologie-Größe wurden nicht gefunden. Der Median für *Filler* lag bei 7 bp für die LB und bei 12 bp für die RB. Ausgeprägte Mikrohomologie sowie *Filler* sind also ein wesentliches Merkmal von T-DNA-Insertionen. Es scheint geringfügig größere *Filler* an der RB zu geben, sonst unterscheiden sich die beiden *Border* hinsichtlich der untersuchten Merkmale nicht.

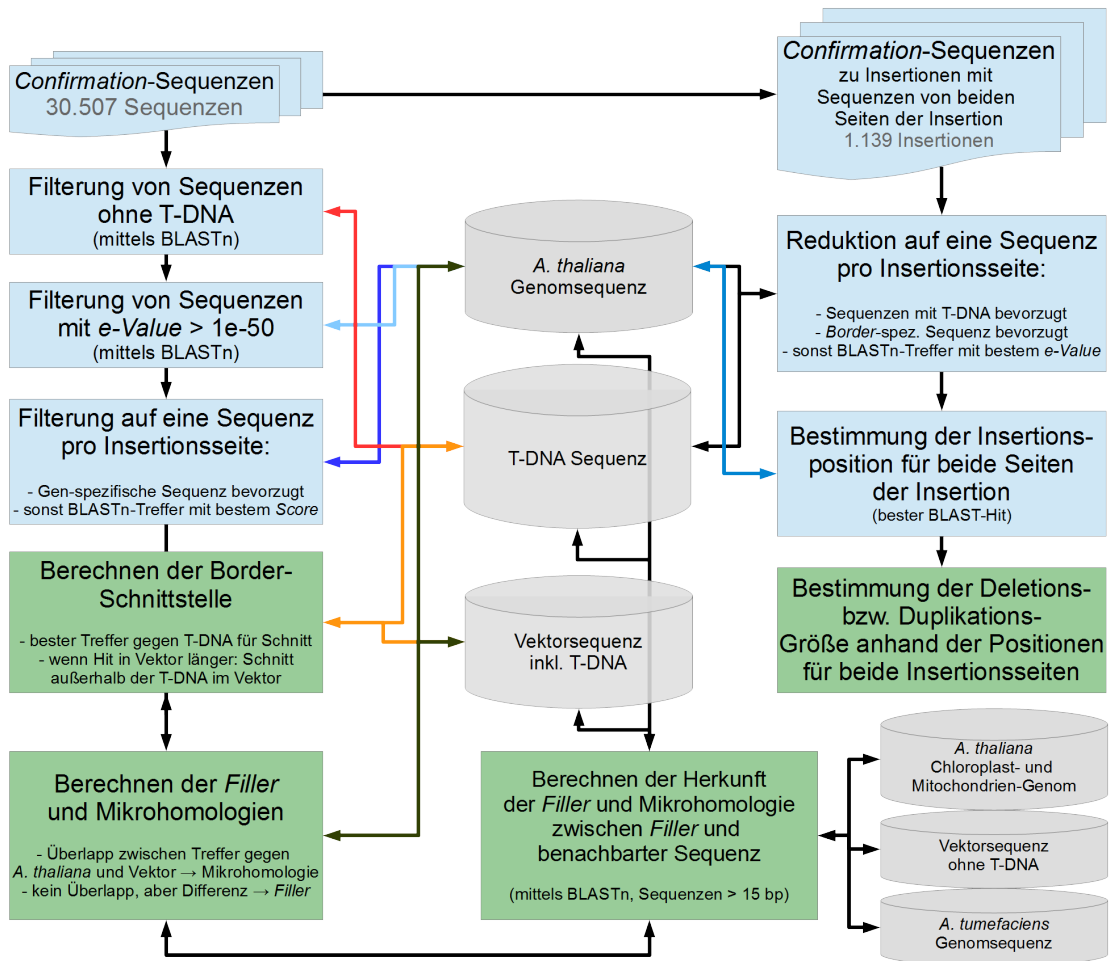


Abbildung 4.8: Pipeline zur Auswertung von Charakteristika am Übergang von T-DNA in genomische DNA. Als Basis für die Analysen dienen *Confirmation*-Sequenzen. Diese werden zunächst gefiltert, sodass nur eine Sequenz pro Flanke einer Insertion übrig bleibt, die ein Stück T-DNA enthält sowie einen maximalen $e\text{-Value}$ bei einer BLAST-Analyse gegen das *A. thaliana*-Kerngenom aufweist. Anschließend kann anhand der BLAST-Resultate gegen den kompletten Vektor inklusive T-DNA die *Border*-Schnittstelle festgelegt werden. Zur Auswertung von Mikrohomologien und *Filler* werden zusätzlich Treffer gegen das *A. thaliana*-Kerngenom herangezogen. Größere *Filler* werden schließlich auf ihren Ursprung untersucht und Mikrohomologie zu benachbarter Sequenz berechnet. Ein Teil der *Confirmation*-Sequenzen wird zur Auswertung von Deletion und Duplikationen an der Insertionsstelle benutzt. Es wird eine Sequenz pro Flanke der Insertion festgelegt und anhand der Differenz ihrer beiden Insertionspositionen die Größe der Deletion oder Duplikation berechnet.

4.4.3 Analyse der *Filler*

Für eine genauere Untersuchung der *Filler* wurden diejenigen mit einer Länge von mindestens 15 bp ausgewählt und mittels BLAST auf ihre Herkunft mit möglichen Quellen verglichen. Diese Quellen umfassen das *A. thaliana*-Kerngenom, das Mitochondrien-Genom, das Chloroplasten-Genom, das *A. tumefaciens*-Genom inklusive Plasmid, den GABI-Kat Vektor ohne die T-DNA sowie die GABI-Kat T-DNA. Die Analyse von 1.508 Sequenzen ergab, dass der Großteil (94,9 %) Homologie zum Kerngenom von *A. thaliana* aufweist. Dabei ist mit 36,3 % ein signifikant hoher Anteil auf demselben Chromosom zu finden anstelle der im Schnitt zu erwartenden 20 % bei 5 Chromosomen. Der überwiegende Rest weist Homologie zu den Randbereichen der T-DNA auf, sowie einige wenige Treffer im Genom von *A. tumefaciens*. Unter den Treffern auf demselben Chromosom gibt es zudem eine Tendenz zu Sequenzteilen in direkter genomischer Nachbarschaft (≤ 1 kbp) der Insertionsstelle. Alle *Filler*, deren Herkunft identifiziert wurde, konnten in einem weiteren Schritt auf Mikrohomologie zu benachbarten Sequenzteilen untersucht werden. Sowohl bei der Suche nach Mikrohomologie zur T-DNA als auch bei der Suche nach Mikrohomologie zum *A. thaliana*-Kerngenom ergab sich ein ähnliches Bild. In 22,8 % der untersuchten Fälle fand sich mindestens 1 bp große Mikrohomologie zur T-DNA und in 25,8 % Mikrohomologie zum Kerngenom von *A. thaliana*. Eine 1 bp große Mikrohomologie ist zwar in ca. 25 % der Fälle aufgrund von zufälliger Sequenzähnlichkeit zu erwarten, größere Mikrohomologien sind aber auch hier deutlich häufiger zu beobachten, als dass sie ein Ergebnis zufälliger Sequenzähnlichkeit sein könnten.

4.4.4 Analyse von Deletionen und Duplikationen an der Insertionsstelle

Insgesamt wurden 1.319 individuelle Insertionen an beiden Übergängen der T-DNA in das *A. thaliana*-Kerngenom untersucht. Insertionspositionen für jede Flanke der Insertion wurden anhand je einer *Confirmation*-Sequenz bestimmt und die Deletions- oder Duplikationsgröße (D) bestimmt durch die Differenz der Insertionsposition auf der strangabwärts liegenden Seite der Insertion ($Pos_{downstream}$) und der strangaufwärts liegenden ($Pos_{upstream}$):

$$D = Pos_{downstream} - Pos_{upstream} - 1 \quad (4.5)$$

Ein negativer Wert stellt somit eine Duplikation dar, ein positiver Wert eine Deletion. Nur 12 der untersuchten Insertionen zeigten keine Veränderung der Sequenz an der Insertionsstelle in Form einer Deletion oder Duplikation. Deletionen wurden in 90,4 % aller Fälle beobachtet mit einem Median von 19 bp. In 6,7 % der Fälle lag eine Duplikation vor, mit einem Median von 4 bp.

Eine Integration ohne Veränderung der Insertionsstelle ist also sehr selten, normalerweise führt sie zu Deletionen von 1-50 bp oder kleineren Duplikationen. Nur in

5 % der Fälle konnten größere Veränderungen (größer 100 bp für Duplikationen, größer 200 bp für Deletionen) beobachtet werden.

4.4.5 Größere Deletionen und Inversionen

Es gab 31 Fälle von Deletionen mit mehr als 1 kbp, 8 von ihnen waren größer als 5 kbp. Bei diesen Insertionen sind oft mehrere Gene betroffen. Um auszuschließen, dass es sich um zwei unabhängige Insertionen oder um eine Translokation handelt, wurden zwei von ihnen, Linie 144F03 und 478B05, näher analysiert.

In Linie 144F03 wurde eine Deletion von 6.154 bp identifiziert. Zur Validierung dieser Deletion wurde DNA von homozygoten T3-Pflanzen extrahiert und mittels Illumina MiSeq sequenziert. Nach *Trimming* der Sequenzier-Rohdaten mit Trimmomatic Version 0.27 [Bolger et al., 2014] und *Mapping* gegen die *A. thaliana*-Genomsequenz mittels BWA [Li and Durbin, 2009] ergab sich kein treffender *Read* im mutmaßlich deletierten Bereich. Die Deletion konnte damit bestätigt und der deletierte Bereich nicht durch eine Translokation erklärt werden.

In Linie 478B05 ist unter anderem das Gen *At3g51240*, das für eine Flavonon-3-Hydroxylase (F3H) kodiert, von einer 5.888 bp großen Deletion betroffen. F3H ist in der Biosynthese von Flavonoiden beteiligt und *Knockouts* zeigen den sogenannten *transparent testa* (*tt*) Phänotyp, der unter anderem an hellen Samen aufgrund eines reduzierten Gehalts von braunen Proanthocyanen erkennbar ist [Debeaujon et al., 2003]. Die homozygoten T3-Pflanzen von 478B05 zeigen alle phänotypischen Merkmale, die auch bei einem anderen bekannten *Knockout* von F3H beobachtet werden können. Die Deletion konnte somit ebenfalls bestätigt werden.

Zusätzlich zu großen Deletionen konnten in 10 Linien anhand der *Confirmation*-Sequenzen Inversionen auf einer Flanke der Insertion nachgewiesen werden.

Die Größe reicht dabei von 52 bp bis 465 bp. Größere Inversionen sind jedoch nicht auszuschließen, da diese durch die Bestätigungs-Prozedur in GABI-Kat nicht gefunden werden können.

4.4.6 Vergleich mit Daten von SALK-Linien

Um zu überprüfen, ob die beobachteten Charakteristika ausschließlich in GABI-Kat zu finden sind, wurden einige SALK-Linien in die Analyse aufgenommen. Deletionen und Duplikationen wurden an 69 individuellen SALK-Insertionen untersucht. Mit einem Median von 19,5 bp unterscheidet sich diese Veränderung an der Insertionsstelle nur geringfügig von der in GABI-Kat beobachteten. Auch kleinere Duplikationen traten bei einzelnen SALK-Insertionen auf.

Des Weiteren wurden 880 Sequenzen auf Mikrohomologie, *Filler* und durch Vir-Proteine innerhalb von *A. tumefaciens* und Abbau in *A. thaliana* verursachte Schnittstellen an den Enden der T-DNA untersucht. Davon stammen 852 Sequenzen von der LB und 28 von der RB. An der RB unterscheidet sich die Schnittposition im Median um 3,5 bp von der erwarteten Position an der Erkennungssequenz. Dies entspricht dem Wert der Analyse von GABI-Kat Sequenzen. An der LB gab

es allerdings eine unerwartet große Differenz von 17 bp. Es lassen sich jedoch zwei Maxima in der Verteilung ausmachen, eines an der erwarteten Schnittstelle und eines etwa 20 bp davon entfernt, was darauf schließen lässt, dass hier eine weitere mögliche Schnittstelle liegt.

Mikrohomologie zwischen T-DNA und *A. thaliana*-Kerngenom wurde in 47 % der untersuchten Sequenzen gefunden mit einem Median von 3 bp (3 bp für LB und 2 bp für RB). *Filler* hatten einen Median von 8 bp (7 bp für LB und 17 bp für RB). Eine genauere Untersuchung von 109 Fillern ergab, dass alle Homologie zum *A. thaliana*-Kerngenom aufweisen und vermehrt (32 %) vom gleichen Chromosom wie die Insertion stammen. Auch diese Ergebnisse stimmen zumindest für die LB, wo eine ausreichende Menge von Sequenzen ausgewertet wurden, mit den GABI-Kat Werten überein.

Zusammenfassende Diskussion

Im Rahmen dieser Arbeit wurden verschiedene Methoden zur Qualitätsverbesserung der Insertionslinienpopulation GABI-Kat und der zugehörigen Dokumentation entwickelt. Dies umfasste zunächst die Re-Annotation der FSTs unter Berücksichtigung der aktuellsten verfügbaren Genom-Annotation von *A. thaliana* (TAIRv10). Damit einher ging eine verbesserte Berechnung der FST-basierten Insertionspositions-Vorhersage unter Berücksichtigung der Länge des FSTs und der Lesbarkeit von T-DNA Sequenz an dessen Beginn. Eine Analyse von bestätigten Insertionen zeigte, dass die verbesserte Vorhersage die Abweichung von der bestätigten Position im Schnitt mehr als halbierte und damit eine deutliche Genauigkeitssteigerung darstellt. Eine falsch vorhergesagte Insertionsposition kann zur Folge haben, dass eine Insertion nicht als Gentreffer klassifiziert wird und in Einzelfällen dazu führen, dass deren Bestätigung fehlschlägt. Eine potentiell interessante Insertion geht in beiden Fällen verloren.

Zur exakteren Bewertung von einzelnen Insertionen wurden neue Kriterien für Gentreffer eingeführt, die es dem Benutzer einfacher machen, geeignete Insertionen zu identifizieren und zu beurteilen.

Außerdem wurden Maßnahmen getroffen, um die Bestätigung einzelner Insertionen wahrscheinlicher zu machen. Dies umfasste die Einführung von Kontaminationsgruppen, die dazu dienen, unter mehreren gleichen Insertions-Vorhersagen in verschiedenen Linien diejenige zu identifizieren, die eine echte Insertion darstellt. Die Verbliebenen gehen auf Kontaminationen untereinander beispielsweise aufgrund von Samen-Vertauschungen, Unsauberkeiten bei der PCR oder DNA-Aufreinigung zurück. Zusätzlich wurden inkorrekte FST-Zuweisungen in einzelnen Blöcken (entsprechen 96-er zur Aufzucht der Linien) korrigiert.

Insgesamt führten diese Optimierungen zu einer Steigerung der Bestätigungsrate, die angibt, wie viele der bearbeiteten Insertionen letztendlich experimentell bestätigt werden konnten, von 78 % auf 84 %. Dies dokumentiert, dass die getroffenen Maßnahmen entscheidend dazu beitrugen, Arbeiten in der Insertionslinienpopulation verlässlicher zu machen und dadurch weniger Allele verworfen werden müssen. Nach der Re-Annotation zeigte sich zudem, dass noch viele unbestätigte Insertionen in GABI-Kat vorhanden sind, die einzigartige Insertionen für Gene darstellen, die in keiner anderen Kollektion vorkommen. Insbesondere diese Allele stellen eine wertvolle Ressource für die weitere Forschung dar. Die Verbesserung der Bestätigungsrate trägt insgesamt dazu bei, diese Ressource zu erhalten.

Das Genom von *A. thaliana* weist aufgrund verschiedener Evolutionsereignisse viele paraloge Bereiche auf. Diese stellen unter Anderem ein Problem dar, wenn *Knockouts* in Genen untersucht werden sollen, die in mehreren Kopien vorliegen. Wie häufig das zu einem Problem werden kann, wurde durch eine Suche nach Paaren von Genen, die für genau zwei ähnliche Proteinen kodieren, untersucht. Diese Analyse ergab, dass unter den angewandten Ähnlichkeitskriterien (siehe Kapitel 4.2) etwa 20 % der Protein-kodierenden Gene in solchen Paaren vorliegen. Etwa die Hälfte hatte mindestens ein zweites Homolog im Genom von *A. thaliana*. Gegenüber publizierten Zahlen ist dies relativ wenig [Armisen et al., 2008]. Die gewählten Parameter waren also relativ streng, daher umfassen die berechneten Genpaare nur sehr ähnliche Proteine. Innerhalb solcher Gruppen homologer Gene mit hoher Ähnlichkeit untereinander wird eine *Knockout*-Mutante in einem einzelnen Gen wahrscheinlich nicht zu einer Veränderung des Phänotyps führen. Die Resultate dieser Analyse sind über die DUPLOdb Webseite zugänglich und bieten eine wertvolle Informationsquelle für Forscher zur Überprüfung, ob der *Knockout* mehrerer Gene bei der Untersuchung eines einzelnen Gens erforderlich ist.

Die identifizierten Genpaare können durch Kreuzung zweier Mutanten verhältnismäßig schnell zu interessanten Phänotypen führen. In GABI-DUPLO wurden vielversprechende Kandidaten für solche Doppelmutanten identifiziert, in 200 Fällen generiert und näher untersucht. Direkt sichtbare Phänotypen zeigten sich in 13 Mutanten (6,5 %). In einer Studie mit 4.000 Einzelmutanten zeigten nur 140 (3,5 %) einen sichtbaren Phänotyp [Kuromori et al., 2006]. Die Ergebnisse bestätigen somit die Ausgangs-Hypothese, dass in Doppelmutanten wesentlich häufiger ein interessanter Phänotyp beobachtet werden kann.

In acht Einzelfällen zeigten sich abweichende Phänotypen von bereits bekannten Doppelmutanten. Die Abweichungen zeigten sich in weniger starken Phänotypen als bisher bekannt, wie nicht vorhandene Embryo-Letalität oder in stärkeren Ausprägungen wie deutlich kleineren Pflanzen. Grund dafür ist vermutlich, dass andere *Knockout*-Linien verwendet wurden. Diese Ergebnisse deuten darauf hin, dass es auch bei Doppelmutanten wichtig ist, immer mehrere verschiedene *Knockout*-Mutanten zu vergleichen, um einen verlässlichen Phänotyp ableiten zu können.

Des Weiteren wurde untersucht, ob Phänotypen in *Knockout*-Mutanten für Genpaare gehäuft dort auftreten, wo starke Sequenzähnlichkeit besteht oder die Expression beider Gene stark korreliert. Dies war in beiden Annahmen der Fall. Derartige Paare sind also vielversprechende Kandidaten für weitere Analysen. Eine Korrelation von Sequenzähnlichkeit und Expression gab es allerdings nicht. Dies deutet darauf hin, dass Promotor- und kodierende Regionen unterschiedlichen Selektionseinflüssen ausgeliefert sind oder sich unabhängig voneinander diversifizieren.

Vermehrte Arbeiten mit Insertionen in paralogen Bereichen erforderten neue bioinformatische Lösungen für die daraus entstehenden Probleme im GABI-Kat LIMS, wie Uneindeutigkeiten bei den FST-Vorhersagen und Schwierigkeiten mit der Generierung einzigartiger Primer. Zwei neue Methoden wurden für diese Fragestellungen etabliert - paraloge Gruppen und ein optimiertes Primerdesign.

Um Mehrdeutigkeiten bei Insertions-Vorhersagen auf Basis eines einzelnen FSTs untersuchen zu können, war es zunächst erforderlich, diese aus den BLAST-Ergebnissen abzuleiten. Dafür wurden die einzelnen Treffer in verschiedene Kategorien eingeteilt und nur eine begrenzte Anzahl von Vorhersagen erlaubt (siehe Kapitel 4.3). Wenn eine Linie mehrere Insertionen enthält, lassen sich in einem FST oft die Sequenzen beider Insertionen finden, da die kürzere FST-Sequenz ein stärkeres Signal während der Sequenzierung erzeugt. Durch die Erweiterung des Datenmodells war es möglich, diese zusätzlichen Insertionen Nutzern zugänglich zu machen. Die Anzahl der vorhergesagten Insertionen stieg durch diese Maßnahme von 91.383 auf 102.494, also um mehr als 10 %. Es ist zu erwarten, dass sich ein Großteil dieser Insertionen bestätigen lässt, da sie nicht auf paralogen Bereichen beruhen.

Zusätzlich wurden Vorhersagen abgespeichert, die zu den aus zusammengesetzten FSTs abgeleiteten Insertionen eine starke Ähnlichkeit aufweisen. Diese wurden in paralogen Gruppen vereint. Nach Bestellung einer solchen Insertion wurden vielfach alle Vorhersagen untersucht und in der Regel nur eine Insertion bestätigt. Viele dieser paralogen Gruppen konnten auf diese Weise aufgeklärt werden. Es zeigte sich, dass in der überwiegenden Anzahl von Fällen die Insertion mit dem besten BLAST-Ergebnis auch die bestätigte Insertion war. In ca. 15 % der untersuchten Gruppen konnte jedoch eine gleich gute oder knapp schlechtere Vorhersage bestätigt werden. Ohne die Untersuchung aller Insertionen in diesen Gruppen wäre die korrekte Insertion nicht identifizierbar gewesen.

Insbesondere bei der Untersuchung von paralogen Gruppen ist es entscheidend, dass die für eine klärende PCR verwendeten Primer eine Unterscheidung ermöglichen, also möglichst nur an einer Stelle im Genom ein Amplimer bilden können. Da der innerhalb der T-DNA liegende Primer durch die jeweilige *Border* festgelegt ist, kann nur der Gen-spezifische Primer so generiert werden, dass er einzigartig im Genom ist. Zu diesem Zweck wurde ein Werkzeug zum Primerdesign entwickelt und implementiert. Es zeichnet sich dadurch aus, die Anzahl möglicher Hybridisierungen für einen Primer, der in einer definierten Zielregion liegen soll, zu minimieren. Einzigartig sollen dabei die letzten 12 Basen des Primers sein, da sich gezeigt hat, dass diese bereits ausreichen, um ein Produkt zu erzeugen. Im Unterschied zu beispielsweise PrimerBlast [Ye et al., 2012], bei dem Primerpaare daraufhin untersucht werden, wie viele Amplimere sie im Genom bilden können, wird jeder einzelne Primer untersucht. Die Fokussierung der Analysen auf die Einzigartigkeit des Gen-spezifischen Primers ist besonders bei der Untersuchung von T-DNA Insertionen von entscheidender Bedeutung.

Das entwickelte Werkzeug wurde neben der internen Benutzung auch der Öffentlichkeit über die GABI-Kat Webseite zugänglich gemacht. Ein besonderes Augenmerk wurde darauf gelegt, das Programm möglichst einfach und intuitiv benutzbar zu machen. Seit der Freischaltung der Funktionalität im September 2014 wurde es bereits 1.957 mal verwendet (Stand Ende März 2015). Dies unterstreicht die Nützlichkeit für andere Forscher bei der Untersuchung von T-DNA Insertionen aus beliebigen Kollektionen.

Ein Indikator für die Zuverlässigkeit des gesamten Prozesses in GABI-Kat und die weiteren Optimierungen ist erneut die Bestätigungsrate, die bereits durch Maßnahmen wie verbesserte Insertionsstellen-Vorhersagen, Aufklärung von Kontaminationen sowie der Untersuchung von falsch zugewiesenen FSTs in einzelnen Blöcken auf insgesamt 84 % gesteigert werden konnte. Ein optimiertes Primerdesign und die verbesserte Bearbeitung von Insertionen in paralogen Bereichen brachten eine weitere Steigerung auf aktuell ca. 85,5 %. Bei Betrachtung von ausschließlich besten Vorhersagen eines FSTs, was vergleichbarer mit dem Ausgangswert von 78 % ist (da nur eine Vorhersage verfügbar war, als dieser Wert berechnet wurde), liegt dieser Wert bei ca. 88 %. Dies stellt seit Beginn der Optimierungen insgesamt eine Steigerung von etwa 10 % dar. Das bedeutet ein fast halbiertes Risiko für einen Benutzer, dass die Bestätigung einer für ihn wichtigen Insertion fehlschlägt. Insgesamt unterstreicht die Gesamtverbesserung der Bestätigungsrate die Steigerung der Datenqualität in GABI-Kat.

Neben der Etablierung von neuen Methoden zur verbesserten Arbeit in einer FST-basierten Kollektion von Insertionsmutanten wurden spezifische Merkmale der durch *A. tumefaciens* vermittelten Transformation untersucht. Als Basis für diese Analysen dienten die *Confirmation*-Sequenzen, die einen einzigartigen Datensatz darstellen, da GABI-Kat die einzige Kollektion ist, die Insertionen auf Basis von Sequenzen bestätigt.

Die Analyse ergab, dass Insertionen im Allgemeinen sehr heterogen sind. Eine Insertion ohne Veränderung einer der Ausgangssequenzen (T-DNA oder *A. thaliana*-Kerngenom) findet äußerst selten statt. Normalerweise wird eine Integration begleitet von kleineren Deletionen bis ca. 50 bp in der Wirts-Sequenz mit einem fließenden Übergang zu kleineren Duplikationen an der Insertionsstelle. Auch an den Grenzen der T-DNA finden häufig Veränderungen der Ausgangssequenz statt. Oft ist der integrierte Teil kürzer, als anhand der Schnittstelle innerhalb der *Border*-Sequenz zu erwarten wäre. Zusätzlich ist ein direkter Übergang von T-DNA in genomische DNA selten. Am Übergang zwischen den beiden Sequenzteilen finden sich gehäuft Mikrohomologie oder *Filler*. Diese *Filler* wiederum stammen vor allem aus dem Kerngenom von *A. thaliana* mit einer Tendenz zu Sequenzen nahe der Insertionsstelle. Des Weiteren lässt sich auch zwischen *Fillern* und der benachbarten DNA (T-DNA oder *A. thaliana*-Kerngenom) öfter ein Bereich mit Mikrohomologie beobachten, als zufällige Sequenzähnlichkeiten erwarten lassen.

All diese Charakteristika unterstützen die These, dass Doppelstrangbruch-Reparaturmechanismen des Wirts für die Integration der T-DNA maßgeblich sind (siehe Kapitel 2.3). Im Einzelstrangbruch-Reparatur Modell sowie im Mikrohomologie-abhängigen Modell zur T-DNA Integration lassen sich die vielfach beobachteten Duplikationen an der Insertionsstelle nicht erklären. Die Ähnlichkeit der beobachteten Merkmale an beiden *Borders* deutet darauf hin, dass im Gegensatz zum Mikrohomologie-abhängigen Modell derselbe Mechanismus für beide Enden der T-DNA verwendet wird. Alle beobachteten Charakteristika belegen dagegen die Theorie des Doppelstrangbruch-Reparatur Modells. Kürzere Mikro-

homologien werden bei Reparaturen durch *Non-Homologous End-Joining* (NHEJ) beobachtet und auch Deletionen an der Reparaturstelle lassen sich durch diesen Mechanismus erklären [Salomon and Puchta, 1998]. Die beobachteten Mikrohomologien sind bis zu 20 bp groß, was darauf schließen lässt, dass nicht nur NHEJ eine Rolle spielt, sondern auch andere verfügbare Mechanismen wie *Microhomology-Mediated End-Joining* (MMEJ), bei dem größere Bereiche von Mikrohomologie genutzt werden [McVey and Lee, 2008]. Dass keine größeren Mikrohomologien beobachtet wurden konnten, liegt an der mangelnden Sequenzhomologie der T-DNA zum *A. thaliana*-Kerngenom. Ein Sequenzvergleich der Sequenzen nahe der *Border* ergab keine längeren Homologien zum Genom von *A. thaliana*.

Die Anwesenheit von *Fillern* und ihre Homologie zu im Kerngenom verfügbaren Sequenzen deutet darauf hin, dass auch homologe Reparaturmechanismen des Wirts eine Rolle spielen. *Synthesis-Dependant Strand-Annealing* (SDSA) kann auch ohne vorhandene Sequenzhomologie dazu führen, dass ein offener Doppelstrangbruch anhand einer verfügbaren Vorlage verlängert und ein *Filler* an der Reparaturstelle erzeugt wird.

Kürzlich wurde eine Studie publiziert, in der keine großen Unterschiede bei der Transformationseffizienz in Mutanten beobachtet werden konnten, die *Knockouts* in Genen haben, die für Schlüsselproteine des NHEJ-Reparaturmechanismus kodieren [Park et al., 2015]. Allerdings wurden die Insertionen selbst nicht untersucht. Es kann spekuliert werden, dass Mikrohomologie zwischen T-DNA und *A. thaliana*-Kerngenom in diesen *Knockouts* größer ist als im transformierten Wildtyp, wenn davon ausgegangen wird, dass andere Reparaturmechanismen zum Einsatz kommen. Dies wird unterstützt durch ältere Resultate einer Studie, bei der CHO-Zellen mit einem *Knockout* in Ku80 untersucht wurden. Die reparierten Doppelstrangbrüche wiesen bei diesen Zellen größere Bereiche an Mikrohomologie, sowie größere Deletionen auf [Feldmann et al., 2000].

Zusätzlich zu den regelmäßig auftretenden Insertions-Charakteristika wurden einige wenige besondere Insertionen untersucht. Für zwei von ihnen konnten größere Deletionen nachgewiesen werden (je ca. 6kb), sowie 10 Inversionen an der Insertionsstelle. Diese Fälle verdeutlichen, dass eine eingehende Analyse einer T-DNA Insertion unerlässlich ist, bevor Rückschlüsse aufgrund eines Phänotyps gezogen werden, der auf solche Veränderungen des Kerngenoms zurückzuführen sein könnte. Insbesondere sollten stets beide Seiten einer Insertion untersucht werden.

Nach fast 14 Jahren, in denen durch *A. tumefaciens* vermittelte Transformation erstellte *Knockout*-Mutanten in vielen Forschungsarbeiten Anwendung fanden, scheint es, dass neuere Methoden zur zielgerichteten Mutagenese von Pflanzengenomen eine vielversprechende Alternative bieten können. In den letzten Jahren haben Zinkfinger-Nukleasen (ZFNs) sowie *transcription activator-like effector nucleases* (TALENs) [Gaj et al., 2013] stetig mehr Anwendungen gefunden.

Beide erlauben die definierte Erzeugung von Doppelstrangbrüchen, bei deren Reparatur durch NHEJ oft aufgrund von Deletionen oder Insertionen eine Leserasterverschiebung entsteht, die das betroffene Gen ausschalten.

Eine Zinkfinger-Domäne besteht aus etwa 30 Aminosäuren. Einige davon binden je nach Zusammenstellung mehr oder weniger spezifisch 3 Nukleotide. Durch Entwicklung von Proteinen mit 3 oder mehr Zinkfingern ist eine recht spezifische Erkennung einer Zielsequenz von bis zu 18 bp möglich. Das Wissen hierzu ist soweit fortgeschritten, dass für jede denkbare Sequenz ein passendes Zinkfinger-Protein synthetisiert werden kann [Gaj et al., 2013].

TALE-Proteine sind ebenfalls in der Lage spezifische DNA-Sequenzen zu binden. Sie enthalten eine 33-35 bp lange *Repeat*-Domäne mit zwei variablen Aminosäuren, die festlegen, welches Nukleotid gebunden wird. Durch Kopplung dieser *Repeat*-Domänen kann eine Ziel-Sequenz festgelegt werden, die mit einem T beginnen muss. Sie sind technisch schwieriger herzustellen als Zinkfinger-Proteine, bieten dafür jedoch mehr Flexibilität, da sie nicht auf Basen-Triplets festgelegt sind.

Durch Fusion mit einer Nuklease können zielgenau Doppelstrangbrüche erzeugt werden, aber auch Transkriptions-Aktivatoren oder Rekombinasen können eingesetzt werden. Durch zusätzliches Einbringen eines Donor-Plasmids mit für die Schnittstelle homologen Armen, können unter Ausnutzung von homologer Reparatur auch gezielt Sequenzen in den erzeugten Doppelstrangbruch eingebracht werden. Die einfachste denkbare Anwendung ist jedoch die zielgerichtete Generierung von *Knockout*-Mutanten.

Als Alternative zu TALENs und ZFNs hat das CRISPR/Cas-System in letzter Zeit vermehrt für Aufmerksamkeit gesorgt [Cong et al., 2013]. Es funktioniert durch eine RNA-gesteuerte DNA-Endonuklease (Cas-Protein). Durch Einbringung kurzer (zur Zielsequenz komplementären) DNA-Abschnitte in den CRISPR-Locus werden diese im Zielorganismus transkribiert und in kleine CRISPR-RNA (crRNA) Moleküle prozessiert, die an transaktivierende crRNAs (tracrRNA) binden. Diese bindet wiederum an das Cas-Protein und zusammen kann eine spezifische Sequenz im Genom erkannt und geschnitten werden. Das CRISPR/Cas-System bietet also die Möglichkeit, die mit TALENs und ZFNs relativ umständlichen Möglichkeiten zur Genom-Editierung einfacher und effizienter zu gestalten und neue Chancen für die Pflanzenforschung und die Generierung von *Knockout*-Mutanten in *A. thaliana* zu eröffnen [Li et al., 2013].

Insbesondere für einzelne Gene, für die es zum aktuellen Zeitpunkt in keiner Kollektion eine Insertionslinie gibt, ist ein zufallsbasierter Mechanismus wie bei der T-DNA Integration nicht mehr zielführend. ZFNs, TALENs oder das CRISPR/Cas-System können in Zukunft eine sinnvolle Alternative bieten.

Existierende, durch *A. tumefaciens* erstellte *Knockout*-Mutanten stellen dagegen immer noch den schnellsten Weg dar, eine Mutante in einem interessanten Gen zu erhalten. Die in dieser Arbeit vorgestellten Methoden haben entscheidend dazu beigetragen, Insertionslinien in GABI-Kat zu einem verlässlicheren Werkzeug zu machen und zeigen Möglichkeiten auf, verschiedene Probleme bei der Arbeit mit einer Insertionslinien-Kollektion zu lösen. Der aus den bestätigten T-DNA Insertionen gewonnene Wissensgewinn trug zu einer Vielzahl von Publikationen bei, in denen GABI-Kat-Linien verwendet wurden. Beispielsweise konnte die Rolle von zwei Genen aufgeklärt werden, die bei der Reaktivierung von durch Methylierung

unterdrückten Genen eine Rolle spielen [Moissiard et al., 2012]. Die Beteiligung mehrerer Gene der AUX/LAX-Familie am Auxin-Import konnte nachgewiesen werden [Peret et al., 2012] oder auch die Rolle zweier Regulatoren der Resistenz gegen die Grauschimmelfäule [Moffat et al., 2012]. Bis Anfang des Jahres 2014 waren 607 Publikationen bekannt, die *Knockout*-Linien von GABI-Kat verwendet haben. Da in dieser Zahl nur solche erfasst werden, die eine GABI-Kat Publikation zitieren, ist die eigentliche Anzahl vermutlich deutlich größer. Alle bestätigten Linien werden weiterhin über das NASC angeboten und die konstant gebliebene Nachfrage danach belegt, dass weltweit immer noch großer Bedarf nach GABI-Kat-Linien besteht und diese auch in Zukunft zu wertvollen Erkenntnisgewinnen führen können.

Die in dieser Arbeit dargestellten Analysen von T-DNA Insertionsstellen in *A. thaliana* trugen dazu bei, den Integrationsmechanismus besser zu verstehen. Alle publizierten Ergebnisse zusammengenommen, kann relativ sicher davon ausgegangen werden, dass die Integration mit Hilfe von verschiedenen Reparaturmechanismen des Wirts-Organismus stattfindet. Insbesondere konnte in dieser Arbeit gezeigt werden, dass größere Bereiche von Mikrohomologie und Deletionen relativ häufig auftreten. Dies lässt vermuten, dass MMEJ eine größere Rolle spielt als zunächst angenommen und eine Integration überwiegend, aber nicht ausschließlich unter Ausnutzung des NHEJ-Mechanismus funktioniert.

Literaturverzeichnis

- Albright, L. M., Huala, E., and Ausubel, F. M. (1989). Prokaryotic signal transduction mediated by sensor and regulator protein pairs. *Annual Review of Genetics*, 23(1):311–336.
- Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., Shinn, P., Stevenson, D. K., Zimmerman, J., Barajas, P., Cheuk, R., et al. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, 301(5633):653–657.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.
- Anand, A., Krichevsky, A., Schornack, S., Lahaye, T., Tzfira, T., Tang, Y., Citovsky, V., and Mysore, K. S. (2007). *Arabidopsis* VIRE2 INTERACTING PROTEIN2 is required for *Agrobacterium* T-DNA integration in plants. *The Plant Cell*, 19(5):1695–1708.
- Armisen, D., Lecharny, A., and Aubourg, S. (2008). Unique genes in plants: specificities and conserved features throughout evolution. *BMC Evolutionary Biology*, 8(1):280.
- Bevan, M. (1984). Binary *Agrobacterium* vectors for plant transformation. *Nucleic Acids Research*, 12(22):8711–8721.
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.
- Boulton, S. J. and Jackson, S. P. (1996). Identification of a *Saccharomyces cerevisiae* Ku80 homologue: roles in DNA double strand break rejoining and in telomeric maintenance. *Nucleic Acids Research*, 24(23):4639–4648.
- Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422(6930):433–438.
- Britt, A. B. and May, G. D. (2003). Re-engineering plant gene targeting. *Trends in Plant Science*, 8(2):90–95.
- Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biology*, 4(1):10.

- Chilton, M.-D., Drummond, M. H., Merlo, D. J., Sciaky, D., Montoya, A. L., Gordon, M. P., and Nester, E. W. (1977). Stable incorporation of plasmid DNA into higher plant cells: the molecular basis of crown gall tumorigenesis. *Cell*, 11(2):263–271.
- Chilton, M.-D. M. and Que, Q. (2003). Targeted integration of T-DNA into the tobacco genome at double-stranded breaks: new insights on the mechanism of T-DNA integration. *Plant Physiology*, 133(3):956–965.
- Citovsky, V., Wong, M. L., and Zambryski, P. (1989). Cooperative interaction of *Agrobacterium* VirE2 protein with single-stranded DNA: implications for the T-DNA transfer process. *Proceedings of the National Academy of Sciences*, 86(4):1193–1197.
- Clough, S. J. and Bent, A. F. (1998). Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *The Plant Journal*, 16(6):735–743.
- Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339(6121):819–823.
- Daley, J. M. and Wilson, T. E. (2005). Rejoining of DNA double-strand breaks as a function of overhang length. *Molecular and Cellular Biology*, 25(3):896–906.
- De Buck, S., Jacobs, A., Van Montagu, M., and Depicker, A. (1999). The DNA sequences of T-DNA junctions suggest that complex T-DNA loci are formed by a recombination process resembling T-DNA integration. *The Plant Journal*, 20(3):295–304.
- Debeaujon, I., Nesi, N., Perez, P., Devic, M., Grandjean, O., Caboche, M., and Lepiniec, L. (2003). Proanthocyanidin-accumulating cells in *Arabidopsis* testa: regulation of differentiation and role in seed development. *The Plant Cell Online*, 15(11):2514–2531.
- Duerrenberger, F., Cramer, A., Hohn, B., and Koukolikova-Nicola, Z. (1989). Covalently bound VirD2 protein of *Agrobacterium tumefaciens* protects the T-DNA from exonucleolytic degradation. *Proceedings of the National Academy of Sciences*, 86(23):9154–9158.
- DUPLOdb (2015). <http://www.gabi-kat.de/duplo.html>.
- Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8(3):186–194.
- Feldmann, E., Schmiemann, V., Goedecke, W., Reichenberger, S., and Pfeiffer, P. (2000). DNA double-strand break repair in cell-free extracts from Ku80-deficient

- cells: implications for Ku serving as an alignment factor in non-homologous DNA end joining. *Nucleic Acids Research*, 28(13):2585–2596.
- GABI-Kat Webseite (2015). <http://www.gabi-kat.de>.
- Gaj, T., Gersbach, C. A., and Barbas III, C. F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*, 31(7):397–405.
- Galkin, V. E., Wu, Y., Zhang, X.-P., Qian, X., He, Y., Yu, X., Heyer, W.-D., Luo, Y., and Egelman, E. H. (2006). The Rad51/RadA N-terminal domain activates nucleoprotein filament ATPase activity. *Structure*, 14(6):983–992.
- Gelvin, S. B. (2003). *Agrobacterium*-mediated plant transformation: the biology behind the gene-jockeying tool. *Microbiology and Molecular Biology Reviews*, 67(1):16–37.
- Gelvin, S. B. (2010). Plant proteins involved in *Agrobacterium*-mediated genetic transformation. *Annual Review of Phytopathology*, 48:45–68.
- Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., Goldman, B. S., Cao, Y., Askenazi, M., Halling, C., et al. (2001). Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science*, 294(5550):2323–2328.
- Gorbunova, V. and Levy, A. A. (1999). How plants make ends meet: DNA double-strand break repair. *Trends in Plant Science*, 4(7):263–269.
- Guerineau, F., Brooks, L., Meadows, J., Lucy, A., Robinson, C., and Mullineaux, P. (1990). Sulfonamide resistance gene for plant transformation. *Plant Molecular Biology*, 15(1):127–136.
- Haas, B. J., Wortman, J. R., Ronning, C. M., Hannick, L. I., Smith, R. K., Maiti, R., Chan, A. P., Yu, C., Farzad, M., Wu, D., et al. (2005). Complete reannotation of the *Arabidopsis* genome: methods, tools, protocols and the final release. *BMC Biology*, 3(1):7.
- Hartwell, L., Hood, L., and Goldberg, M. L. (2008). *Genetics: from genes to genomes*. Granite Hill Publishers.
- Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919.
- Hernalsteens, J.-P., Van Vliet, F., De Beuckeleer, M., Depicker, A., Engler, G., Lemmers, M., Holsters, M., Van Montagu, M., and Schell, J. (1980). The *Agrobacterium tumefaciens* Ti plasmid as a host vector system for introducing foreign DNA in plant cells. *Nature*, 287:654–656.

- Hoeijmakers, J. H. (2001). Genome maintenance mechanisms for preventing cancer. *Nature*, 411(6835):366–374.
- Hughes, A. L., Friedman, R., Ekollu, V., and Rose, J. R. (2003). Non-random association of transposable elements with duplicated genomic blocks in *Arabidopsis thaliana*. *Molecular Phylogenetics and Evolution*, 29(3):410–416.
- Intrieri, M. C. and Buiatti, M. (2001). The horizontal transfer of *Agrobacterium rhizogenes* genes and the evolution of the genus *Nicotiana*. *Molecular Phylogenetics and Evolution*, 20(1):100–110.
- Jasper, F., Koncz, C., Schell, J., and Steinbiss, H.-H. (1994). *Agrobacterium* T-strand production in vitro: sequence-specific cleavage and 5' protection of single-stranded DNA templates by purified VirD2 protein. *Proceedings of the National Academy of Sciences*, 91(2):694–698.
- Jin, S., Roitsch, T., Ankenbauer, R., Gordon, M., and Nester, E. (1990). The VirA protein of *Agrobacterium tumefaciens* is autophosphorylated and is essential for vir gene regulation. *Journal of Bacteriology*, 172(2):525–530.
- Kim, S.-I. and Gelvin, S. B. (2007). Genome-wide analysis of *Agrobacterium* T-DNA integration sites in the *Arabidopsis* genome generated under non-selective conditions. *The Plant Journal*, 51(5):779–791.
- Kloetgen, A. (2011). Analyse paraloger Gruppen in der Insertionslinienpopulation GABI-Kat, Bachelorarbeit, Universität Bielefeld.
- Krebs, J. E., Lewin, B., Goldstein, E. S., and Kilpatrick, S. T. (2014). *Lewin's Genes XI*. Jones & Bartlett Publishers.
- Krizekova, L. and Hroudá, M. (1998). Direct repeats of T-DNA integrated in tobacco chromosome: characterization of junction regions. *The Plant Journal*, 16(6):673–680.
- Krysan, P. J., Young, J. C., and Sussman, M. R. (1999). T-DNA as an insertional mutagen in *Arabidopsis*. *The Plant Cell*, 11(12):2283–2290.
- Kuromori, T., Wada, T., Kamiya, A., Yuguchi, M., Yokouchi, T., Imura, Y., Takabe, H., Sakurai, T., Akiyama, K., Hirayama, T., et al. (2006). A trial of phenome analysis using 4000 Ds-insertional mutants in gene-coding regions of *Arabidopsis*. *The Plant Journal*, 47(4):640–651.
- Lacroix, B., Vaidya, M., Tzfira, T., and Citovsky, V. (2005). The VirE3 protein of *Agrobacterium* mimics a host cell function required for plant genetic transformation. *The EMBO Journal*, 24(2):428–437.
- Laibach, F. (1943). *Arabidopsis thaliana* als Objekt für genetische und entwicklungsphysiologische Untersuchungen. *Bot. Archiv*, 44:439–455.

- Lee, Y.-W., Jin, S., Sim, W.-S., and Nester, E. W. (1995). Genetic evidence for direct sensing of phenolic compounds by the VirA protein of *Agrobacterium tumefaciens*. *Proceedings of the National Academy of Sciences*, 92(26):12245–12249.
- Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760.
- Li, J., Krichevsky, A., Vaidya, M., Tzfira, T., and Citovsky, V. (2005). Uncoupling of the functions of the *Arabidopsis* VIP1 protein in transient and stable plant genetic transformation by *Agrobacterium*. *Proceedings of the National Academy of Sciences*, 102(16):5733–5738.
- Li, J.-F., Norville, J. E., Aach, J., McCormack, M., Zhang, D., Bush, J., Church, G. M., and Sheen, J. (2013). Multiplex and homologous recombination-mediated genome editing in *Arabidopsis* and *Nicotiana benthamiana* using guide RNA and Cas9. *Nature Biotechnology*, 31(8):688–691.
- Li, Y., Rosso, M. G., Uelker, B., and Weisshaar, B. (2006). Analysis of T-DNA insertion site distribution patterns in *Arabidopsis thaliana* reveals special features of genes without insertions. *Genomics*, 87(5):645–652.
- Li, Y., Rosso, M. G., Viehoveer, P., and Weisshaar, B. (2007). GABI-Kat SimpleSearch: an *Arabidopsis thaliana* T-DNA mutant database with detailed information for confirmed insertions. *Nucleic Acids Research*, 35(Database issue):D874–D878.
- Liang, F., Romanienko, P. J., Weaver, D. T., Jeggo, P. A., and Jasin, M. (1996). Chromosomal double-strand break repair in Ku80-deficient cells. *Proceedings of the National Academy of Sciences*, 93(17):8929–8933.
- Liang, Z. and Tzfira, T. (2013). *In vivo* formation of double-stranded T-DNA molecules by T-strand priming. *Nature Communications*, 4(2253).
- Liu, Y., Maskell, D. L., and Schmidt, B. (2009). CUDASW++: optimizing Smith-Waterman sequence database searches for CUDA-enabled graphics processing units. *BMC Research Notes*, 2(1):73.
- Liu, Y.-G., Mitsukawa, N., Oosumi, T., and Whittier, R. F. (1995). Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. *The Plant Journal*, 8(3):457–463.
- Magori, S. and Citovsky, V. (2011). *Agrobacterium* counteracts host-induced degradation of its effector F-box protein. *Science Signaling*, 4(195):ra69.
- McVey, M. and Lee, S. E. (2008). MMEJ repair of double-strand breaks (directors cut): deleted sequences and alternative endings. *Trends in Genetics*, 24(11):529–538.

- Mezard, C. (2006). Meiotic recombination hotspots in plants. *Biochemical Society Transactions*, 34(4):531–534.
- Moffat, C. S., Ingle, R. A., Wathugala, D. L., Saunders, N. J., Knight, H., and Knight, M. R. (2012). ERF5 and ERF6 play redundant roles as positive regulators of JA/Et-mediated defense against *Botrytis cinerea* in *Arabidopsis*. *PLoS One*, 7(4):e35995.
- Moissiard, G., Cokus, S. J., Cary, J., Feng, S., Billi, A. C., Stroud, H., Husmann, D., Zhan, Y., Lajoie, B. R., McCord, R. P., et al. (2012). MORC family AT-Pases required for heterochromatin condensation and gene silencing. *Science*, 336(6087):1448–1451.
- Mysore, K. S., Nam, J., and Gelvin, S. B. (2000). An *Arabidopsis* histone H2A mutant is deficient in *Agrobacterium* T-DNA integration. *Proceedings of the National Academy of Sciences*, 97(2):948–953.
- Nam, J., Mysore, K., Zheng, C., Knue, M., Matthysse, A., and Gelvin, S. (1999). Identification of T-DNA tagged *Arabidopsis* mutants that are resistant to transformation by *Agrobacterium*. *Molecular and General Genetics*, 261(3):429–438.
- Nam, J., Mysore, K. S., and Gelvin, S. B. (1998). *Agrobacterium tumefaciens* transformation of the radiation hypersensitive *Arabidopsis thaliana* mutants *wh1* and *rad5*. *Molecular Plant-Microbe Interactions*, 11(11):1136–1141.
- Ophel, K. and Kerr, A. (1990). *Agrobacterium vitis* sp. nov. for strains of *Agrobacterium biovar 3* from grapevines. *International Journal of Systematic Bacteriology*, 40(3):236–241.
- Pansegrau, W., Schoumacher, F., Hohn, B., and Lanka, E. (1993). Site-specific cleavage and joining of single-stranded DNA by VirD2 protein of *Agrobacterium tumefaciens* Ti plasmids: analogy to bacterial conjugation. *Proceedings of the National Academy of Sciences*, 90(24):11538–11542.
- Park, S.-Y., Vaghchhipawala, Z., Vasudevan, B., Lee, L.-Y., Shen, Y., Singer, K., Waterworth, W. M., Zhang, Z. J., West, C. E., Mysore, K. S., et al. (2015). *Agrobacterium* T-DNA integration into the plant genome can occur without the activity of key non-homologous end-joining proteins. *The Plant Journal*, 81(6):934–946.
- Peret, B., Swarup, K., Ferguson, A., Seth, M., Yang, Y., Dhondt, S., James, N., Casimiro, I., Perry, P., Syed, A., et al. (2012). AUX/LAX genes encode a family of auxin influx transporters that perform distinct functions during *Arabidopsis* development. *The Plant Cell*, 24(7):2874–2885.
- PGDD (2015). <http://chibba.agtec.uga.edu/duplication/>.
- Ray, A. and Langer, M. (2002). Homologous recombination: ends as the means. *Trends in Plant Science*, 7(10):435–440.

- Rolloos, M., Dohmen, M. H., Hooykaas, P. J., and Zaal, B. J. (2014). Involvement of Rad52 in T-DNA circle formation during *Agrobacterium tumefaciens*-mediated transformation of *Saccharomyces cerevisiae*. *Molecular Microbiology*, 91(6):1240–1251.
- Rosso, M. G., Li, Y., Strizhov, N., Reiss, B., Dekker, K., and Weisshaar, B. (2003). An *Arabidopsis thaliana* T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Molecular Biology*, 53(1-2):247–259.
- Roth, D. B. and Wilson, J. H. (1986). Nonhomologous recombination in mammalian cells: role for short sequence homologies in the joining reaction. *Molecular and Cellular Biology*, 6(12):4295–4304.
- Salomon, S. and Puchta, H. (1998). Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells. *The EMBO Journal*, 17(20):6086–6095.
- Sanford, J. C., Klein, T. M., Wolf, E. D., and Allen, N. (1987). Delivery of substances into cells and tissues using a particle bombardment process. *Particulate Science and Technology*, 5(1):27–37.
- Schell, J., Van Montagu, M., De Beuckeleer, M., De Block, M., Depicker, A., De Wilde, M., Engler, G., Genetello, C., Hernalsteens, J.-P., Holsters, M., et al. (1979). Interactions and DNA transfer between *Agrobacterium tumefaciens*, the Ti-plasmid and the plant host. *Proceedings of the Royal Society of London*, 204(1155):251–266.
- Schrammeijer, B., Risseeuw, E., Pansegrau, W., Regensburg-Tu nk, T. J., Crosby, W. L., and Hooykaas, P. J. (2001). Interaction of the virulence protein VirF of *Agrobacterium tumefaciens* with plant homologs of the yeast Skp1 protein. *Current Biology*, 11(4):258–262.
- Sessions, A., Burke, E., Presting, G., Aux, G., McElver, J., Patton, D., Dietrich, B., Ho, P., Bacwaden, J., Ko, C., et al. (2002). A high-throughput *Arabidopsis* reverse genetics system. *The Plant Cell*, 14(12):2985–2994.
- Shi, Y., Lee, L.-Y., and Gelvin, S. B. (2014). Is VIP1 important for *Agrobacterium*-mediated transformation? *The Plant Journal*, 79(5):848–860.
- Singer, K., Shibolet, Y. M., Li, J., and Tzfira, T. (2012). Formation of complex extrachromosomal T-DNA structures in *Agrobacterium tumefaciens*-infected plants. *Plant Physiology*, 160(1):511–522.
- Smith, E. F. and Townsend, C. O. (1907). A plant-tumor of bacterial origin. *Science*, 25(643):671–673.

- Staden, R. (1996). The Staden sequence analysis package. *Molecular Biotechnology*, 5(3):233–241.
- Stracke, R., Huep, G., and Weisshaar, B. (2010). Use of mutants from T-DNA insertion populations generated by high-throughput screening. *The Handbook of Plant Mutation Screening: Mining of Natural and Induced Alleles*, pages 31–54.
- Strizhov, N., Li, Y., Rosso, M. G., Viehoveer, P., Dekker, K. A., and Weisshaar, B. (2003). High-throughput generation of sequence indexes from T-DNA mutagenized *Arabidopsis thaliana* lines. *Biotechniques*, 35(6):1164–1169.
- Sugawara, N., Ira, G., and Haber, J. E. (2000). DNA length dependence of the single-strand annealing pathway and the role of *Saccharomyces cerevisiae* RAD59 in double-strand break repair. *Molecular and Cellular Biology*, 20(14):5300–5309.
- Sussman, M. R., Amasino, R. M., Young, J. C., Krysan, P. J., and Austin-Phillips, S. (2000). The *Arabidopsis* knockout facility at the University of Wisconsin–Madison. *Plant Physiology*, 124(4):1465–1467.
- Symington, L. S. (2002). Role of RAD52 epistasis group genes in homologous recombination and double-strand break repair. *Microbiology and Molecular Biology Reviews*, 66(4):630–670.
- Szabados, L., Kovacs, I., Oberschall, A., Abraham, E., Kerekes, I., Zsigmond, L., Nagy, R., Alvarado, M., Krasovskaja, I., Gal, M., et al. (2002). Distribution of 1000 sequenced T-DNA tags in the *Arabidopsis* genome. *The Plant Journal*, 32(2):233–242.
- TAIRv10 (2010). ftp://ftp.arabidopsis.org/home/tair/sequences/whole_chromosomes/.
- Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science*, 320(5875):486–488.
- The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796.
- Thompson, J. D., Gibson, T., Higgins, D. G., et al. (2002). Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics*, 2:2–3.
- Tinland, B. (1996). The integration of T-DNA into plant genomes. *Trends in Plant Science*, 1(6):178–184.
- Tinland, B. and Hohn, B. (1994). Recombination between prokaryotic and eukaryotic DNA: integration of *Agrobacterium tumefaciens* T-DNA into the plant genome. *Genetic Engineering*, 17:209–229.

- Tinland, B., Hohn, B., and Puchta, H. (1994). *Agrobacterium tumefaciens* transfers single-stranded transferred DNA (T-DNA) into the plant cell nucleus. *Proceedings of the National Academy of Sciences*, 91(17):8000–8004.
- Tzfira, T., Frankman, L. R., Vaidya, M., and Citovsky, V. (2003). Site-specific integration of *Agrobacterium tumefaciens* T-DNA via double-stranded intermediates. *Plant Physiology*, 133(3):1011–1023.
- Tzfira, T., Li, J., Lacroix, B., and Citovsky, V. (2004a). *Agrobacterium* T-DNA integration: molecules and models. *Trends in Genetics*, 20(8):375–383.
- Tzfira, T., Vaidya, M., and Citovsky, V. (2001). VIP1, an *Arabidopsis* protein that interacts with *Agrobacterium* VirE2, is involved in VirE2 nuclear import and *Agrobacterium* infectivity. *The EMBO Journal*, 20(13):3596–3607.
- Tzfira, T., Vaidya, M., and Citovsky, V. (2002). Increasing plant susceptibility to *Agrobacterium* infection by overexpression of the *Arabidopsis* nuclear protein VIP1. *Proceedings of the National Academy of Sciences*, 99(16):10435–10440.
- Tzfira, T., Vaidya, M., and Citovsky, V. (2004b). Involvement of targeted proteolysis in plant genetic transformation by *Agrobacterium*. *Nature*, 431(7004):87–92.
- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., and Rozen, S. G. (2012). Primer3 - new capabilities and interfaces. *Nucleic Acids Research*, 40(15):e115.
- Vaghchhipawala, Z. E., Vasudevan, B., Lee, S., Morsy, M. R., and Mysore, K. S. (2012). *Agrobacterium* may delay plant nonhomologous end-joining DNA repair via XRCC4 to favor T-DNA integration. *The Plant Cell*, 24(10):4110–4123.
- van Attikum, H., Bundock, P., and Hooykaas, P. J. (2001). Non-homologous end-joining proteins are required for *Agrobacterium* T-DNA integration. *The EMBO Journal*, 20(22):6550–6558.
- van Attikum, H. and Hooykaas, P. J. (2003). Genetic requirements for the targeted integration of *Agrobacterium* T-DNA in *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 31(3):826–832.
- Van Larebeke, N., Engler, G., Holsters, M., Van den Elsacker, S., Zaenen, I., Schilperoort, R., and Schell, J. (1974). Large plasmid in *Agrobacterium tumefaciens* essential for crown gall-inducing ability. *Nature*, 252(5479):169–170.
- Velten, J., Velten, L., Hain, R., and Schell, J. (1984). Isolation of a dual plant promoter fragment from the Ti plasmid of *Agrobacterium tumefaciens*. *The EMBO Journal*, 3(12):2723.

- Vergunst, A. C., van Lier, M. C., den Dulk-Ras, A., Stüve, T. A. G., Ouwehand, A., and Hooykaas, P. J. (2005). Positive charge is an important feature of the C-terminal transport signal of the VirB/D4-translocated proteins of *Agrobacterium*. *Proceedings of the National Academy of Sciences*, 102(3):832–837.
- Wortman, J. R., Haas, B. J., Hannick, L. I., Smith, R. K., Maiti, R., Ronning, C. M., Chan, A. P., Yu, C., Ayele, M., Whitelaw, C. A., et al. (2003). Annotation of the *Arabidopsis* genome. *Plant Physiology*, 132(2):461–468.
- Wyman, C. and Kanaar, R. (2006). DNA double-strand break repair: all's well that ends well. *Annu. Rev. Genet.*, 40:363–383.
- Wyman, C., Ristic, D., and Kanaar, R. (2004). Homologous recombination-mediated double-strand break repair. *DNA Repair*, 3(8):827–833.
- Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, 13(1):134.
- Yoshiyama, K. O., Sakaguchi, K., and Kimura, S. (2013). DNA damage response in plants: conserved and variable response compared to animals. *Biology*, 2(4):1338–1356.
- Zambryski, P., Joos, H., Genetello, C., Leemans, J., Van Montagu, M., and Schell, J. (1983). Ti plasmid vector for the introduction of DNA into plant cells without alteration of their normal regeneration capacity. *The EMBO Journal*, 2(12):2143.
- Zekic, T. (2012). Verbesserung des Primerdesigns in der Insertionslinienpopulation GABI-Kat, Bachelorarbeit, Universität Bielefeld.
- Ziemienowicz, A., Tinland, B., Bryant, J., Gloeckler, V., and Hohn, B. (2000). Plant Enzymes but not *Agrobacterium* VirD2 mediate T-DNA ligation *in vitro*. *Molecular and Cellular Biology*, 20(17):6317–6322.

Abkürzungsverzeichnis

ABRC	Arabidopsis Biological Resource Center
BAC	<i>Bacterial Artificial Chromosome</i>
cM	Centimorgan
EMS	Ethylmethansulfonat
EST	<i>Expressed Sequence Tag</i>
FST	<i>Flanking Sequence Tag</i>
HR	Homologe Reparatur
LB	<i>Left Border</i>
LIMS	Labor Informations Management System
MMEJ	<i>Microhomology-Mediated End-Joining</i>
NASC	Nottingham Arabidopsis Stock Centre
NHEJ	<i>Non-Homologous End-Joining</i>
PCR	<i>Polymerase Chain Reaction</i>
RB	<i>Right Border</i>
SDSA	<i>Synthesis-Dependant Strand-Annealing</i>
SSA	<i>Single-Strand-Annealing</i>
Sul	Sulfadiazin
TAIR	<i>The Arabidopsis Information Resource</i>
TIGR	<i>The Institute for Genomic Research</i>
TE	Transponierbare Elemente
UTR	<i>untranslated region</i>

Publikation 1:

GABI-Kat SimpleSearch: New features of the
Arabidopsis thaliana T-DNA mutant database

GABI-Kat SimpleSearch: new features of the *Arabidopsis thaliana* T-DNA mutant database

Nils Kleinboelting, Gunnar Huep, Andreas Kloetgen, Prisca Viehoveer and Bernd Weisshaar*

Center for Biotechnology (CeBiTec), Bielefeld University, Universitaetsstrasse 25, D-33615 Bielefeld, Germany

Received September 14, 2011; Revised October 24, 2011; Accepted October 25, 2011

ABSTRACT

T-DNA insertion mutants are very valuable for reverse genetics in *Arabidopsis thaliana*. Several projects have generated large sequence-indexed collections of T-DNA insertion lines, of which GABI-Kat is the second largest resource worldwide. User access to the collection and its Flanking Sequence Tags (FSTs) is provided by the front end SimpleSearch (<http://www.GABI-Kat.de>). Several significant improvements have been implemented recently. The database now relies on the TAIRv10 genome sequence and annotation dataset. All FSTs have been newly mapped using an optimized procedure that leads to improved accuracy of insertion site predictions. A fraction of the collection with weak FST yield was re-analysed by generating new FSTs. Along with newly found predictions for older sequences about 20 000 new FSTs were included in the database. Information about groups of FSTs pointing to the same insertion site that is found in several lines but is real only in a single line are included, and many problematic FST-to-line links have been corrected using new wet-lab data. SimpleSearch currently contains data from ~71 000 lines with predicted insertions covering 62.5% of the 27 206 nuclear protein coding genes, and offers insertion allele-specific data from 9545 confirmed lines that are available from the Nottingham Arabidopsis Stock Centre.

INTRODUCTION

Since the genome sequence of the model plant *Arabidopsis thaliana* was completed in the year 2000 (1), the determination of gene function is a key task in the Arabidopsis research community. Insertional mutagenesis approaches have been proven to be a valuable tool for reverse

genetics (2). Several large collections of *A. thaliana* lines containing independent insertions of *Agrobacterium* T-DNA in the plant genome have been established. T-DNA integration in the plant genome results in stable mutations, which may perturb gene functions. An important goal is to saturate the *A. thaliana* genome with T-DNA insertion lines for each (or at least most) of the 27 206 nuclear protein-coding genes that are currently annotated (3). A popular strategy to determine the position of a T-DNA insertion in the genome of a given insertion line is to generate Flanking Sequence Tags (FSTs). Using PCR-based methods, genomic DNA fragments flanking the T-DNA are amplified (4), sequenced and subsequently mapped to the genome. When this is applied to a large collection of lines, the population can easily be searched for mutants of interest.

GABI-Kat is the second largest FST-indexed T-DNA insertion line collection of *A. thaliana*, which is publicly available since 2002 (5). User access to the collection and extensive metadata for the included mutants and alleles is provided by GABI-Kat SimpleSearch, the web interface of the corresponding database (6). The interface can either be used to search the collection for mutant alleles of interest, or more importantly to access different kinds of information about specific GABI-Kat lines. Lines containing insertions of interest can be ordered via a web order form, which is also provided in SimpleSearch. In contrast to other FST databases, GABI-Kat SimpleSearch offers information on confirmation success of lines along with sequences derived from the T-DNA/genome junction of an offspring generation, segregation data and primer information for the respective insertions (7).

Since its availability in 2002, the database as well as the interface has been continuously improved. However, during the last 18 months several significant improvements and extensions have been implemented that are beneficial for users that rely on SimpleSearch when working with GABI-Kat insertion alleles. These enhancements include an extended data set, allow an easier and more comfortable user access, and provide more detailed and more

*To whom correspondence should be addressed. Tel: +49 521 106 8720; Fax: +49 521 106 6423; Email: bernd.weisshaar@uni-bielefeld.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2011. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

reliable meta-information about the insertion alleles in the GABI-Kat collection. In detail, the recent improvements are (i) an update to the most recent *Ath* genome annotation dataset TAIRv10, (ii) an improved insertion site prediction and gene hit definition, (iii) enhanced information about confirmation success and line availability. Together with the correction of problematic FST-to-line links using new experimental data, these enhancements result in an increased overall quality and reliability of the collection and its database.

RESULTS AND DISCUSSION

General database content

The SimpleSearch database contains data about GABI-Kat insertion mutants, the derived FSTs and their mapping to the *A. thaliana* genome sequence. In addition and in contrast to other FST databases like SIGnAL (8) or FLAGdb++ (9), SimpleSearch focuses on metadata concerning GABI-Kat lines and insertion alleles. The database contains data about confirmation attempts for predicted insertion alleles, genetic segregation data of the resistance phenotype provided by the T-DNA allowing an estimation of the number of insertion loci per line, and sequences of successfully confirmed alleles from the offspring generation that allow to determine the real site of insertion much more accurately than the crude FST sequences. Moreover, the user is provided with information about line availability (alleles in dead lines are lost although the respective FST exists in databases), and about insertions that could not be confirmed (failed insertions, which are predicted from FSTs but are not existing in following generations). The updated version of SimpleSearch offers significantly improved data quality due to intensive and ongoing quality management and manual curation (see below). GABI-Kat FSTs are produced by an adaptor-ligation PCR method (4) (see also the Methods and FAQ pages on <http://www.gabi-kat.de/>) and are mapped to the *A. thaliana* genome using BLAST (10). Users can access the FST data, select alleles of interest and place insertion requests. Upon request the GABI-Kat lines are segregated on selective medium, genomic DNA is prepared and the predicted insertion sites are confirmed by PCR and sequencing, using an insertion site specific primer and a T-DNA border primer. The obtained 'confirmation sequences' are again mapped to the genome. If the insertion site deduced from confirmation sequencing matches the position predicted from the respective FST, the insertion is regarded as confirmed.

Update of the database to TAIRv10

Until recently, the SimpleSearch database (7) was based upon the TIGR version 5 annotation dataset of the *A. thaliana* genome. The TIGRv5 annotation consisted of individual BAC sequences and contains an outdated set of gene annotation data (11). SimpleSearch has now been updated to the current genome annotation data, namely TAIR version 10, which is based on pseudochromosomes (ftp://ftp.arabidopsis.org/home/tair/home/tair/Sequences/whole_chromosomes/). In this context all FSTs from the

GABI-Kat collection were newly mapped to the genome sequence, and putative insertion positions were deduced. The increase in the number of successfully mapped FSTs results from using optimized mapping parameters, reduced gaps in the genome sequence and also from the generation of new FSTs for fractions of the GABI-Kat population with formerly low FST yield.

About 20 000 new FSTs have been submitted to EMBL/GenBank. These contribute to a total of now 130 000 FSTs at EMBL/GenBank, which are also accessible via SimpleSearch. The reliability of the insertion site prediction was improved by deducing the insertion position from the annealing site of the T-DNA border-specific primer and the nucleotide positions from the original trace-file of the FST (Figure 1) (12,13). Particularly in cases where no T-DNA sequence is detectable in the FST, the prediction of the insertion positions is more accurate now. This advantage is evident when compared to insertion positions deduced from the called and trimmed sequence only (8). However, deletions at the T-DNA border to genome junction still cause errors in the prediction, which can only be resolved by detailed analysis of the confirmation sequences. Nevertheless, we now explicitly predict a defined pseudochromosome position as insertion position, and not only a locus that is described by a gene code or BAC name.

For a better assessment of the relevance of a predicted insertion allele, we used data from TAIRv10 to further qualify hits with respect to annotated genes. TAIRv10 contains information about the untranslated region of the mRNA (UTR) for the majority of the protein coding genes, as well as information about RNA-coding genes. Our former definition defined a 'gene hit' as an insertion site prediction between 300-bp upstream of the ATG and 300-bp downstream of the stop codon of a gene, whereas a 'CDSi hit' had a predicted insertion between ATG and stop codon (CDSi for coding sequence plus

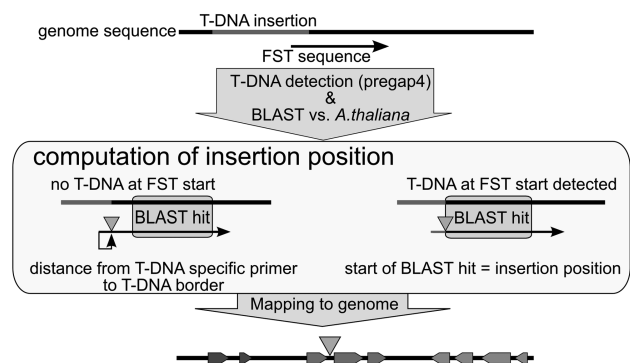


Figure 1. Workflow of the improved insertion site prediction. The insertion position is determined using the best BLAST hit of the FST sequence vs. the *A. thaliana* genome sequence, and the location of the T-DNA within the FST sequence determined by pregap4. If no T-DNA is detected at the start of the FST sequence, the insertion site is located x bases upstream of the BLAST hit, where x is the number of bases before the start of the BLAST hit minus the distance of the T-DNA specific primer to the T-DNA border. Otherwise, the start of the BLAST hit is considered as insertion position.

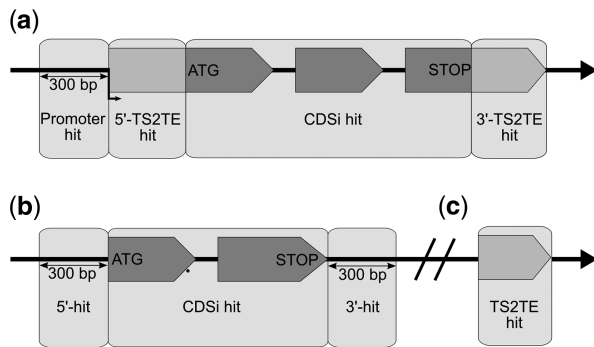


Figure 2. Definition of gene hits at GABI-Kat. (a) For protein-coding genes with annotated UTR-regions in TAIRv10, we differentiate between CDSi hits (insertion position between ATG and STOP), 5'- and 3'-TS2TE hits (insertion position in the 5'- or 3'-UTR) and promoter hit [insertion position up to 300-bp upstream of transcription start (TS)]. (b) If the UTR is not annotated in TAIRv10 (and for pseudogenes), insertion positions 300-bp up or downstream of ATG and STOP are considered as 5'- and 3'-hits. (c) For RNA genes and transposable elements, TS2TE hits are annotated, if the insertion is located between TS and transcript end.

introns). We extended this definition by including hits in the transcribed region of a gene as well as the promoter region (Figure 2). 'TS2TE hits' have an insertion between transcription start and transcript end, and 'promoter hits' have the predicted insertion position within 300-bp upstream of the transcription start. This definition applies for protein- as well as for RNA-coding genes. If no UTR is annotated for a given gene, we still use the old gene hit definition. All hits that are not linked to a gene keep the insertion type 'genome hit' (Figure 2). Due to the fact that the old 'gene hit' definition covered a larger genome area than the TS2TE area of the genome, the total number of insertions that qualify as 'gene hits' went slightly down, although the number of predicted insertions in the database did increase. The updated insertion types allow a more reliable selection of insertion alleles that may be NULL alleles.

Overall, the total number of lines with predicted insertions was increased to 71 235. Table 1 gives a summary of the current data content of the SimpleSearch database.

Increase of quality and accuracy of the database content

Due to the manual handling steps during the generation of the GABI-Kat population, it is an unpleasant fact that some links between FST sequences and insertion lines are wrong. Reasons for these errors can be mix-ups during plant growth in the greenhouse, handling errors during sequence generation, or alike. Incorrectly assigned FST-sequences were detected when the confirmation rate in a set of 96 lines (corresponding to a microtitre plate) was much lower than expected from average values. Based on the knowledge about the FST production workflow, we deduced different types of errors that might have happened and reassigned the FST-sequences to the correct lines after wet-lab validation of the hypothesis. This reassignment of FST-sequences has meanwhile been performed for the majority of problematic microtitre

Table 1. Summary of data in the GABI-Kat SimpleSearch database^a

Data type	Number of entries
FSTs ^b	~133 000
Lines ^b	71 235
with segregation data	15 289
available at NASC	9644
Insertions with predicted insertion position ^c	88 580
analysed with final result ^d	16 081
delivered to individual users	6816
confirmed and available at NASC ^e	9653
Distinct genes covered	21 005
protein coding genes	19 120
ncRNA coding genes	182
pseudogenes	420
transposable element genes	1283
Gene hits available only from GABI-Kat ^f	2114
Confirmed 'GABI-Kat only' hits at NASC	1201
'GABI-Kat only' hits to be addressed ^g	765
Distinct CDSi covered	13 037

^aNumbers as of 15 September 2011.

^bDatabase release version 24 (affects FSTs and lines that are in the database, the data values for the items in the database are updated every 24 h).

^cInsertions are different from lines, because a line can contain several insertions. Example: 011F01, which is confirmed for a genome hit at F26P21 (Chr4) and a TS2TE hit in At5g05180.

^dA final result can be 'confirmed', but also 'failed to confirm' or 'part of a contamination group' are considered.

^eFor each confirmed insertion there are confirmation sequences available which are generated from the amplicon that spans the T-DNA/genome sequence junction.

^fOnly hits that may cause a NULL allele (CDSi hits and hits in the 5'-UTR) are counted. Only lines in the accession Columbia-0 are considered, which is the accession used by the main FST-based insertion line collections.

^gAbout 150 'GABI-Kat only' alleles are either in the queue already and wait for mature T3 seed, or did fail to confirm.

plates and contributed to a significant increase (~3%) of the confirmation rate of insertion alleles from the collection. The corrected FST-to-line connections have been integrated into the database and are available to users through SimpleSearch. It should be noted that SimpleSearch is currently the only source of these corrections. Until other FST databases have been updated, the now detected and corrected errors in the original FST dataset are still 'proliferated' from e.g. the old FST data in sequence databases.

When predictions for very similar insertion sites are found in multiple lines, we combine these lines into 'contamination groups'. The assumption is that the prediction is only true for one of those lines and the others are caused by contaminations that happened during high-throughput DNA-preparation, PCR, or sequencing. If an insertion that is part of a contamination group is ordered by a user, the whole contamination group is examined experimentally. Finally, the correct insertion is delivered to the user and donated to the Nottingham Arabidopsis Stock Centre (NASC). In some cases, several alleles are confirmed with closely linked insertion positions, and in these cases all alleles except one confirmed allele are removed from the 'contamination group'. SimpleSearch contains information about failed insertion confirmations

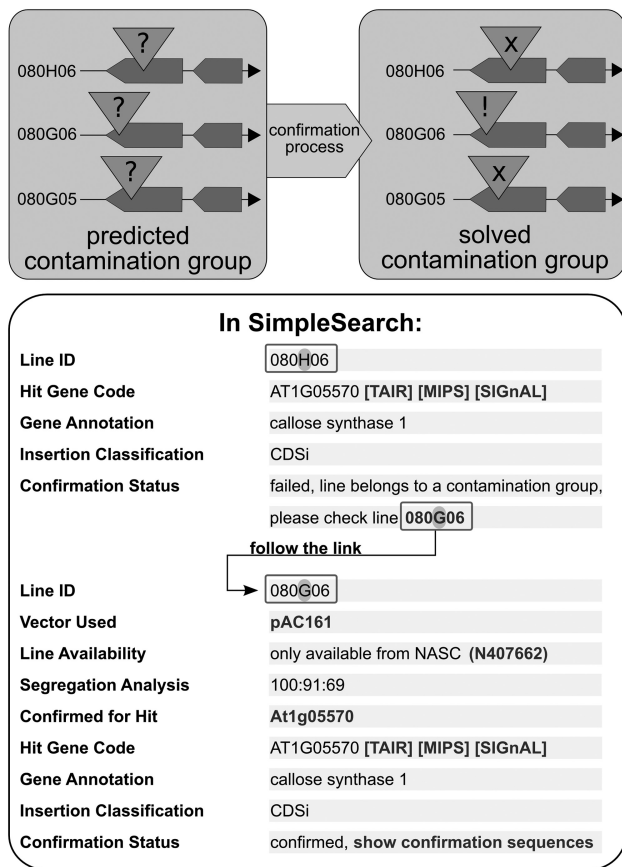


Figure 3. Resolution of contamination groups. A contamination group contains predicted insertions in different lines that share very similar insertion positions (within 50 bp at most). After the confirmation process, only one line is confirmed, the others failed and are considered as contaminations. When searching for insertion alleles in SimpleSearch, the user is guided to confirmed allele if the contamination group is solved.

and leads the user to the confirmed allele within a (resolved) contamination group (Figure 3). Until September 2011 about 1250 possible contamination groups could be solved and the information is available in SimpleSearch.

As a result of various actions addressing data quality, including those mentioned above, the quality of the collection has been improved significantly. This is best quantified with the increase of the confirmation rate from 78% to 84%. Without information from the unique in-house confirmation process that is carried out at GABI-Kat, this big step towards improved reliability would have been impossible.

With a total of 21 005 genes that are covered with insertions [counted are ‘gene hits’ if no UTR is annotated, ‘promoter hits’ and ‘TS2TE hits’ (which include ‘CDSi hits’)], the current database release v24 covers about 4000 more genes with (predicted) insertion alleles than described earlier (7). This also includes genes for non-coding RNAs (ncRNAs), which were not annotated in TIGRv5. Due to the small size of many ncRNA-coding genes, the coverage is fairly low (182 of 1290 for ncRNA

genes), which also affects the statistics of total gene hit coverage. The total coverage of the nuclear protein coding genes with (predicted) insertion alleles has increased from 64% to 70%.

Changes in the web interface

The static part of the GABI-Kat website is realised with an Open Source CMS system. SimpleSearch is embedded into this website, and the dynamic pages were developed in PHP. Information about individual items in SimpleSearch, e.g. a given line or a list of hits in a given gene, can be accessed by a defined URL format (see the help pages of the website). Both the static and the dynamic part look identical to the user. Besides visual improvement, the information content has been extended. When searching for insertions, SimpleSearch now offers an overview of insertion alleles and their availability. Lines already donated to NASC (blue triangle), insertions that failed in the confirmation process and are therefore unavailable (red triangle), and lines available for entering the confirmation process at GABI-Kat (green triangle) are distinguished. In case that a line with a failed confirmation attempt belongs into a solved contamination group, the correctly confirmed line for the insertion is linked in the SimpleSearch interface, which enables the user to access the confirmed line easily.

In addition to the existing options to search for single lines, FSTs, hits in a given gene or by BLAST, we added a search option that lists all (predicted) insertion positions in a position range on the pseudochromosomes.

Perspective

The GABI-Kat resource has served the *A. thaliana* community as a valuable tool for reverse genetics since it was made available to the public in 2002. The demand for GABI-Kat insertions was constantly high since then, which is documented by the number of about 72 000 stock requests at NASC until February 2011. Until September 2011 9644 different confirmed GABI-Kat lines have been donated to the stock centre. In addition to continue to confirm and deliver insertion alleles to users, we are currently addressing about 760 lines with new insertion predictions in genes for which no allele is available in the other main FST-based *A. thaliana* Col-0 insertion line collections.

SimpleSearch offers an easy and well-featured access to data about GABI-Kat lines and confirmed or unconfirmed insertion alleles. To maintain the quality, it is important to constantly curate the database and adopt it to the most recent knowledge about the *A. thaliana* genome, e.g. an upcoming v11 *A. thaliana* genome release. In parallel, we are also working on additional topics, which could be improved. One problem arises from (short) FSTs that are assigned to genes of which paralogous copies exist in the genome. In such cases the insertion position prediction and the assignment to a single locus is error-prone. It is a challenge to represent the ‘paralog problem’ in the database, and to address it experimentally quite some wet-lab work would be required. This example shows that there is still room for further improvement.

However, the current status and the up-to-date data content of the SimpleSearch database have reached a very comprehensive level through the measures described in this article.

ACCESSION NUMBERS

About 20 000 FSTs have been submitted to EMBL/GenBank: FR799760–FR819654.

ACKNOWLEDGEMENTS

The authors thank Yong Li, Mario Rosso, the MPI for Plant Breeding Research and all former co-workers for their contribution to GABI-Kat, and Renate Harder, Helene Schellenberg, Nina Schmidt, Andrea Voigt, Marja-Leena Wilke for technical assistance.

FUNDING

German Federal Ministry of Education and Research (BMBF) in the context of the German plant genomics program GABI (Förder Kennzeichen 0313855). Funding for open access charge: Bielefeld University.

Conflict of interest statement. None declared.

REFERENCES

1. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
2. O'Malley, R.C. and Ecker, J.R. (2010) Linking genotype to phenotype using the *Arabidopsis* unimutant collection. *Plant J.*, **61**, 928–940.
3. Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L. *et al.* (2008) The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
4. Strizhov, N., Li, Y., Rosso, M.G., Viehoveer, P., Dekker, K.A. and Weisshaar, B. (2003) High-throughput generation of sequence indexes from T-DNA mutagenized *Arabidopsis thaliana* lines. *BioTechniques*, **35**, 1164–1168.
5. Rosso, M.G., Li, Y., Strizhov, N., Reiss, B., Dekker, K. and Weisshaar, B. (2003) An *Arabidopsis thaliana* T-DNA mutagenised population (GABI-Kat) for flanking sequence tag based reverse genetics. *Plant Mol. Biol.*, **53**, 247–259.
6. Li, Y., Rosso, M.G., Strizhov, N., Viehoveer, P. and Weisshaar, B. (2003) GABI-Kat SimpleSearch: a flanking sequence tag (FST) database for the identification of T-DNA insertion mutants in *Arabidopsis thaliana*. *Bioinformatics*, **19**, 1441–1442.
7. Li, Y., Rosso, M.G., Viehoveer, P. and Weisshaar, B. (2007) GABI-Kat SimpleSearch: an *Arabidopsis thaliana* T-DNA mutant database with detailed information for confirmed insertions. *Nucleic Acids Res.*, **35**, D874–D878.
8. Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R. *et al.* (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, **301**, 653–657.
9. Samson, F., Brunaud, V., Duchene, S., De Oliveira, Y., Caboche, M., Lecharny, A. and Aubourg, S. (2004) FLAGdb++: a database for the functional analysis of the *Arabidopsis* genome. *Nucleic Acids Res.*, **32**, D347–D350.
10. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. Wortman, J.R., Haas, B.J., Hannick, L.I., Smith, R.K. Jr, Maiti, R., Ronning, C.M., Chan, A.P., Yu, C., Ayele, M., Whitelaw, C.A. *et al.* (2003) Annotation of the *Arabidopsis* genome. *Plant Physiol.*, **132**, 461–468.
12. Bonfield, J.K., Smith, K.F. and Staden, R. (1995) A new DNA sequence assembly program. *Nucleic Acids Res.*, **23**, 4992–4999.
13. Staden, R., Beal, K.F. and Bonfield, J.K. (2000) The Staden package, 1998. *Methods Mol. Biol.*, **132**, 115–130.

Publikation 2:

GABI-DUPLO: a collection of double mutants to overcome genetic redundancy in *Arabidopsis thaliana*

TECHNICAL ADVANCE/RESOURCE

GABI-DUPLO: a collection of double mutants to overcome genetic redundancy in *Arabidopsis thaliana*

Cordelia Bolle¹, Gunnar Huep², Nils Kleinbölting², Georg Haberer³, Klaus Mayer³, Dario Leister^{1,*} and Bernd Weisshaar²¹Lehrstuhl für Molekularbiologie der Pflanzen (Botanik), Department Biologie I, Ludwig-Maximilians-Universität München, Großhaderner Str. 2, D-82152, Planegg-Martinsried, Germany,²Genome Research, Department of Biology, Bielefeld University, 33594 Bielefeld, Germany, and³MIPS, Institute for Bioinformatics and Systems Biology, Helmholtz Center Munich, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany

Received 11 March 2013; revised 28 March 2013; accepted 4 April 2013; published online 10 April 2013.

*For correspondence (e-mail leister@lmu.de).

SUMMARY

Owing to duplication events in its progenitor, more than 90% of the genes in the *Arabidopsis thaliana* genome are members of multigene families. A set of 2108 gene families, each consisting of precisely two unlinked paralogous genes, was identified in the nuclear genome of *A. thaliana* on the basis of sequence similarity. A systematic method for the creation of double knock-out lines for such gene pairs, designated as DUPLO lines, was established and 200 lines are now publicly available. Their initial phenotypic characterisation led to the identification of seven lines with defects that emerge only in the adult stage. A further six lines display seedling lethality and 23 lines were lethal before germination. Another 14 lines are known to show phenotypes under non-standard conditions or at the molecular level. Knock-out of gene pairs with very similar coding sequences or expression profiles is more likely to produce a mutant phenotype than inactivation of gene pairs with dissimilar profiles or sequences. High coding sequence similarity and highly similar expression profiles are only weakly correlated, implying that promoter and coding regions of these gene pairs display different degrees of diversification.

Keywords: *Arabidopsis thaliana*, double mutant, gene duplication, genetic redundancy, segmental duplication, technical advance.

INTRODUCTION

Only 9.7% of the approximately 27 000 protein-coding genes (TAIR 10; <http://www.arabidopsis.org>) in the nuclear genome of the model plant *Arabidopsis thaliana* code for unique proteins; all others have at least one additional homologue (Armisen *et al.*, 2008). Moreover, many of the latter class form larger gene families, with about 40% being accounted for by families with more than five members and some comprising more than 100 members (AGI, 2000; Cannon *et al.*, 2004; Armisen *et al.*, 2008). Genes can be amplified by several mechanisms (for recent reviews see: Van de Peer *et al.*, 2009; Paterson *et al.*, 2010; Rutter *et al.*, 2012). Single-gene duplications such as tandem duplications generate tightly linked copies, whereas duplications resulting from more widespread restructuring processes, such as polyploidization (duplication of the whole genome) or duplications of single chromosomes (aneu-

ploidy) or large chromosomal segments, can lead to the rearrangement of large chromosomal regions and thereby to segmental duplications. Evidence for at least three major large-scale duplication events has been detected in the genome of *A. thaliana*, all of them predating the divergence of *Arabidopsis* species from *Brassica* species about 14.5–20.4 million years ago (Bowers *et al.*, 2003; Jiao *et al.*, 2011). The most recent duplication event encompassed 70–89% of the genes in the genome (Simillion *et al.*, 2002; Blanc *et al.*, 2003; Bowers *et al.*, 2003).

Because duplicated genes are redundant in function immediately after their formation, they are exposed to genetic fractionation processes (AGI, 2000; He and Zhang, 2005). One gene copy may be lost with no deleterious effects. Mutations in one or both gene copies may lead either to a change in gene dosage or to functional divergence

between the two. The latter may involve the gain of a novel function (neofunctionalization) or both gene copies can retain different subsets of the functional properties of the ancestral gene (subfunctionalization) (Lynch and Conery, 2000; Rutter *et al.*, 2012). Duplicated genes can therefore possess overlapping or redundant functions – especially if the duplication event is recent – and various instances of such genetic redundancy have been reported. Among these instances are several photosynthesis-related proteins, such as PsaD, PsaE, plastocyanin and PGRL1, which are encoded by segmentally duplicated genes and for which knock-out of both gene copies uncovers phenotypes not seen in either single mutant (Ihnatowicz *et al.*, 2004, 2007; DalCorso *et al.*, 2008; Pesaresi *et al.*, 2009). Moreover, it has been demonstrated that 22 core cell-cycle genes have evolved from segmental duplications (Vandepoele *et al.*, 2002). Duplications can also lead to the evolution of extended gene families that exhibit various degrees of diversification in sequence and expression pattern (Kolukisaoglu *et al.*, 2002; Meyers *et al.*, 2003; Cannon *et al.*, 2004; Leister, 2004; Remington *et al.*, 2004; Abel *et al.*, 2005; Jiang *et al.*, 2006; Maher *et al.*, 2006; Nakano *et al.*, 2006; Kong *et al.*, 2007; Wang *et al.*, 2009; Liu *et al.*, 2010; Charon *et al.* 2012).

Several collections of sequence-indexed T-DNA insertion mutants have already been constructed for *A. thaliana* (reviewed in Bolle *et al.*, 2011). In total, over 600 000 mutant lines have been generated and flanking sequence tags (FSTs) of over 325 000 T-DNA insertion lines have been mapped to the reference genome sequence (*A. thaliana* accession Columbia 0). More recently, a project was initiated with the intention of generating two independent homozygous T-DNA insertion lines for each nuclear Arabidopsis gene (O'Malley and Ecker, 2010). Such homozygous lines are a prerequisite for the direct assessment of gene function and will enable efficient genome-wide forward genetic screens. Gene silencing by small interfering RNAs or by artificial microRNAs is also an important tool in the study of gene function and can also be applied to the silencing of gene families (Ossowski *et al.*, 2008). However, whereas double mutations caused by T-DNA insertions are stably inherited, leakiness and silencing of RNA-based gene inactivation approaches have been occasionally observed (e.g. Gupta *et al.*, 2002; Weigel *et al.*, 2003). Therefore, lines generated by RNA-based approaches and the DUPLO lines represent complementary approaches to inactivate gene pairs, whereas RNA-based approaches alone are the method of choice to inactivate simultaneously three or more gene copies, including tandem copies.

Here we report on the generation of a collection of double knock-out mutants (DMs) specifically targeted to the members of highly homologous, but genetically unlinked, gene pairs, and derived from sequence-indexed T-DNA mutant lines with a Col-0 genetic background. This collection,

termed 'GABI-DUPLO', is available at the Nottingham Arabidopsis Stock Centre (NASC) and will complement the existing sequence-indexed single-mutant collections by allowing one to bypass the effects of genetic redundancies arising from segmental duplications in systematic forward and reverse genetic screens. An initial phenotypic characterisation of 200 GABI-DUPLO lines is described here; we demonstrate that the probability that a DM line will display a phenotype increases with the degree of similarity in expression pattern or protein sequence between the two genes considered.

RESULTS

Selection of gene pairs for the GABI-DUPLO collection

To unmask the functions of genes that are represented by two copies in the *A. thaliana* genome through the analyses of DM phenotypes, we selected gene pairs that were highly likely to have redundant or overlapping functions. To this end, we employed two selection criteria, namely overall sequence similarity and number of homologous gene copies. In the first step, we selected protein-coding *A. thaliana* genes that have high sequence similarity at the amino acid sequence level ($\geq 60\%$ similar amino acids and $<20\%$ gaps in a global Needleman–Wunsch alignment). This analysis resulted in a set of 26 982 pairwise combinations of closely related genes fulfilling this criterion (Figure 1). To identify true gene pairs, combinations that had at least one additional (third) homologue in the *A. thaliana* genome at the same similarity threshold were eliminated. This action resulted in a set of 2594 gene pairs. Paired genes located ≤ 7.5 Mbp apart on the same chromosome were not considered further as their proximity would have prevented the efficient generation of DMs, so that only 2108 gene pairs were ultimately selected for analysis (Figure 1).

When these 2108 gene families, each made up of exactly two genetically unlinked members, were analysed for their known or predicted molecular function, as well as the biological process they are involved in and the subcellular localization of their products (Figures 2 and S1), the set showed no obvious enrichment for any particular category (relative to the entire *A. thaliana* genome). However, genes coding for proteins whose molecular function is unknown (14 versus 20% within the whole genome), biological context unclear (5% versus 8%) or subcellular localization undetermined (9 versus 19%) are slightly underrepresented among the 2108 gene pairs, whereas the functional categories 'transcription factor activity' and 'DNA/RNA binding' are slightly overrepresented (6/9 versus 3/7%).

The GABI-Kat and SALK collections were analysed for the presence of T-DNA insertion alleles of both members of each of the 2108 gene pairs (Alonso *et al.*, 2003; Kleinboelting *et al.*, 2012). To obtain loss-of-function alleles of the genes of interest, only alleles with a (predicted)

insertion in the region between the start and stop codon (i.e. exon or intron regions and referred to here as 'CDS + intron' or 'CDSi' insertions) were considered. Such insertions were identified for 1294 of the gene pairs (Figure 1). However, because some CDSi insertions were either:

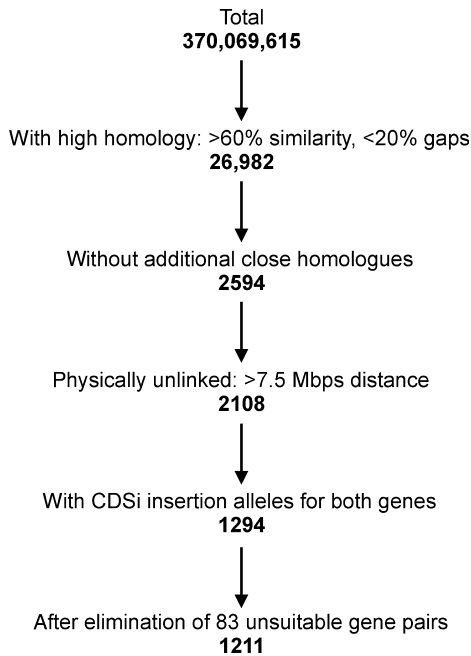


Figure 1. Flow chart that depicts the selection of lines for the DUPLO collection. The total number of all possible pairwise combinations for the 27 206 protein-coding genes in *A. thaliana* is 370 069 615. The stepwise selection of gene pairs is illustrated. Note that the number of gene pairs with CDSi alleles for both genes depends on the availability of insertion alleles and can change either when new alleles become available or when flanking sequence tag (FST)-based predictions are modified, for instance upon re-annotation of genes in the TAIR database. In the final selection step, gene pairs were excluded if the insertion was not found to be located within the CDS (after updating of TAIR to version 10), or the allele could not be confirmed, or if one or both parental lines were known to be lethal. For further details see main text and Experimental Procedures.

(i) unavailable; (ii) no longer annotated as CDSi insertions in version 10 of the TAIR database; (iii) could not be confirmed by polymerase chain reaction (PCR); or (iv) known to be lethal; we finally obtained mutant alleles for 1211 gene pairs that fulfilled all criteria for setting up a DM collection (Figure 1) (for a full list see: <https://www.gabi-kat.de/double-mutant-lists/duplo-gene-pairs.html>). Obviously, this list reflects the availability of insertion alleles at the time of writing and will be expanded as new insertion alleles become available.

Generation of homozygous DM lines

The generation of homozygous DMs usually requires at least 8 months. Parental single-mutant lines were first analysed by PCR and amplicon sequencing for the presence of the T-DNA insertion allele of the gene of interest. These 'confirmation amplicon' sequences were evaluated to deduce the position of the insertion, and only confirmed CDSi alleles were used for further steps in the workflow. If the putative CDSi allele could not be confirmed by sequencing, we tried to obtain a more suitable allele from the GABI-Kat or SALK collection. If this confirmation was not possible, the respective pair was removed from the workflow (see Figure 1). Lines that contained confirmed CDSi alleles were then genotyped using two different primer pairs. The 'genotyping amplicon' targets the wild-type allele, spans the entire insertion site in the mutant line and is generated only if the inserted T-DNA is not present. The 'confirmation amplicon' (see above) targets the insertion allele and is identical to the amplicon sequenced during confirmation (see Experimental Procedures and Figure S2).

In the next step, the parental lines that were homozygous or at least heterozygous for the insertion in each member of a given gene pair were crossed. F1 plants were allowed to self-pollinate if the presence of the two expected insertion alleles could be confirmed by PCR (Figure 3). About 54

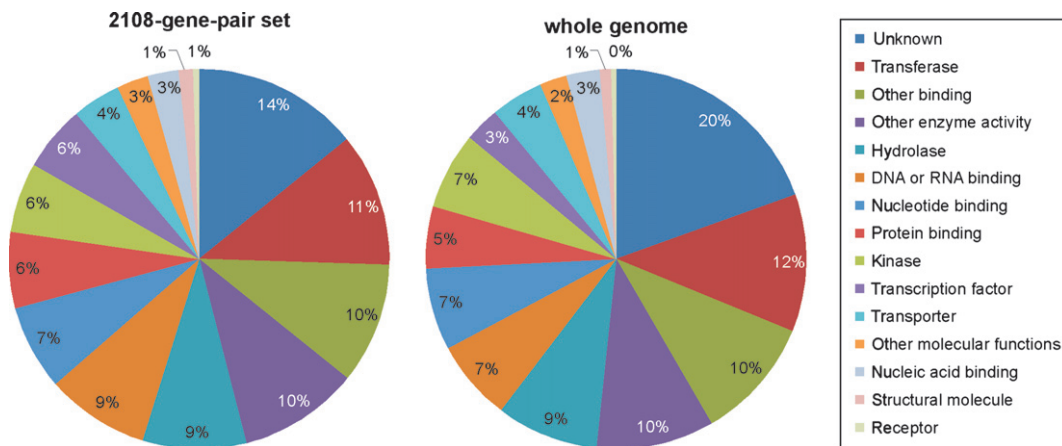


Figure 2. Distribution of the proteins encoded by the 2108 gene pairs according to their predicted molecular function. The molecular functions of the proteins encoded by the 2108 gene pairs (left panel) and by the entire *A. thaliana* genome as control (right panel) were determined and displayed in a pie chart using the Gene Ontology (GO) annotation tool on the TAIR website (<http://www.arabidopsis.org/tools/bulk/go/index.jsp>).

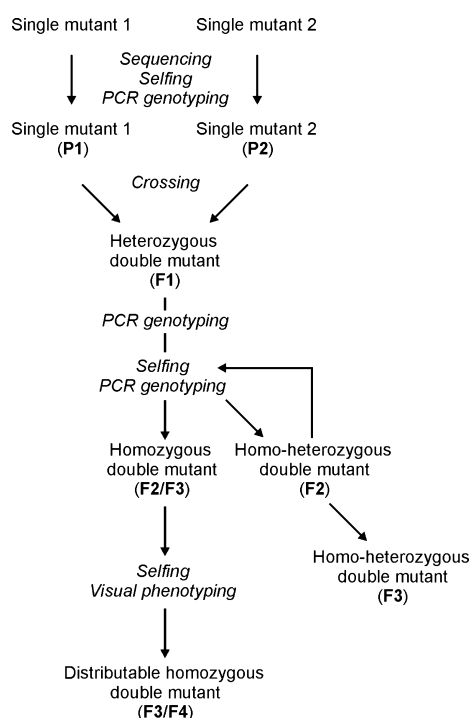


Figure 3. Workflow used for the systematic generation of DMs. Parental lines (P1 and P2) were crossed to generate an F1 generation. After selfing, the F2 generation was genotyped to isolate double knock-out mutants (DMs). If no homozygous DM could be obtained, this step was repeated in the F3 generation. Seeds from homozygous DM lines were bulked in the next generation for distribution. Phenotyping was done in the F3 or F4 generation.

individual F2 plants were then propagated on soil and genotyped by PCR to identify plants that were homozygous for both insertions. Genotyping of the DMs with respect to the insertions in the two duplicated genes was performed by four PCRs, amplifying ‘confirmation’ and ‘genotyping amplicons’ for each of the two mutant alleles (Figure 3).

If no DUPLO line could be identified in the F2 generation, the offspring of an F2 line that was homozygous for one insertion and heterozygous for the other was allowed to self. The segregating F3 generation was propagated on plates of sucrose-containing media to allow heterotrophic growth, and genotyped by PCR (Figure 3). Propagation on sucrose-containing medium not only permitted the rescue of some mutants with defects that interfere with photoautotrophy or normal seedling development but also enabled us to identify non-germinating seeds, which are characteristic for embryo lethality. If segregation of the F3 seeds also failed to yield any homozygous DM, these pairs were classified as ‘homozygous DM absent’ and are likely to be lethal as DMs. In the F3 or F4 generation the genotype of homozygous DMs was rechecked and plants were evaluated phenotypically in the greenhouse.

So far, we have generated 200 DMs that were either homozygous for both mutations, or homozygous for one

mutation and heterozygous for the other, and that are designated in the following as ‘DUPLO lines’ (Table S1). According to the Plant Genome Duplication Database (<http://chibba.agtec.uga.edu/duplication/index/home>) 145 (or 72%) of the 200 gene pairs are located within large segmental duplications (Tang *et al.*, 2008a,b).

DUPLO lines that display mutant phenotypes at the adult stage

During phenotypic evaluation at the F3 or F4 stage we looked for changes in growth and development at different time points. The first evaluation for size and pigmentation took place 14 days after germination (dag). Leaf coloration, rosette size, leaf form and early bolting were evaluated at 30 dag. Once the plants had flowered, their branching pattern, overall height and fertility were monitored. Seven DUPLO lines that deviated from WT plants with respect to chlorophyll content, size or developmental behaviour were identified (Table 1), and this class of mutants was referred to as DMs with an ‘adult phenotype’, because these plants survived to the adult stage and into the reproductive phase. All of these mutants displayed reduced or dwarfish growth, and three mutants had a reduced photosynthetic efficiency. These lines are described in more detail in the following. Unless otherwise noted, the parental single mutants behaved like WT.

The *DUPLO-195* line is a cross of the loss-of-function line of *TON1 recruiting motif 3 (trm3, at1g18620)* and *trm4 (at1g74160)*. Some members of the TRM protein family have been shown to interact with TON1 and are involved in microtubule organisation at the cortex (Drevensek *et al.*, 2012). The DM and *trm4* show late flowering and a reduced height of the inflorescence, whereas *trm3* did not show any visually discernible phenotype (Figure 4a). In addition, DM leaves are smaller and occasionally rolled and display a reduced effective quantum efficiency of photosystem II (Φ_{II}) (Table 2 and Figure 4a).

The *DUPLO-932* line is dwarfed as it forms very small rosettes (about 30% of the WT diameter), and flowering is much delayed under long-day greenhouse conditions (Table 2 and Figure 4b). The molecular function of the two genes mutated (*At2g41770* and *At3g57420*), which are expressed ubiquitously in the plant, is not yet known, and the parental lines show no deviation from WT.

In the *DUPLO-998* line (*at5g18180 at3g03920*), two genes that code for the GAR1 subunit of the H/ACA ribonucleoprotein complex, which has been shown in mammalian systems to have a pseudouridine synthase activity and is important for rRNA folding and ribosome function (Watkins and Bohnsack, 2012), are defective. Although these two genes appear to be the only GAR1 orthologues in Arabidopsis, the DM plants are viable but display reduced growth and later flowering (Table 2 and Figure 4c).

Table 1 List of DUPLO lines displaying a mutant phenotype at the adult or seedling stage. The Pearson correlation coefficient (PCC) from an all-against all comparison of 170 microarray experiments based on the Affymetrix ATH1 platform is provided in the last column

Pair ID	Gene pair	Protein description	Phenotype	PCC
Adult phenotype				
195	<i>At1g18620</i> ; <i>At1g74160</i>	TRM3; TRM4	Smaller plant size, later flowering	0.71
932	<i>At2g41770</i> ; <i>At3g57420</i>	Expressed proteins	Dwarfed, later flowering	0.88
998	<i>At5g18180</i> ; <i>At3g03920</i>	H/ACA ribonucleoprotein complex, subunit Gar1/Naf1 proteins	Smaller plant size, later flowering	0.61
2261	<i>At1g04300</i> ; <i>At5g43560</i>	TRAF-like superfamily proteins	Bushy plant, shorter petioles, stunted	0.92
2268	<i>At3g11910</i> ; <i>At5g06600</i>	UBP13; UBP12	Extremely small, dies often on soil; Ewan <i>et al.</i> , 2011 (conditional phenotype)	0.91
2589	<i>At1g66180</i> ; <i>At5g37540</i>	Eukaryotic aspartyl protease family proteins	Smaller plant size	0.68
2647	<i>At1g51340</i> ; <i>At3g08040</i>	MATE efflux family protein; FRD3, MAN1	Smaller plant size, less chlorophyll	0.61
Seedling lethal				
211	<i>At1g76090</i> ; <i>At1g20330</i>	Sterol methyltransferase 3; CVP1	Seedling extremely small, not viable; Carland <i>et al.</i> , 2010 (smaller and bushy, short siliques, poor seed yield)	0.78
713	<i>At1g76490</i> ; <i>At2g17370</i>	HMGR1; HMGR2	Seedling extremely small, not viable; Suzuki <i>et al.</i> , 2009 (male gametophytes lethal)	0.68
1569	<i>At5g47420</i> ; <i>At4g17420</i>	Tryptophan RNA-binding attenuator protein-like proteins	Seedling extremely small, dies early	0.89
2098	<i>At4g24330</i> ; <i>At5g49945</i>	Proteins of unknown function (DUF1682)	Seedling extremely small, not viable	0.91
2490	<i>At1g23310</i> ; <i>At1g70580</i>	GGT1; GGT2	Seedling extremely small, dies early	0.76
2543	<i>At1g74910</i> ; <i>At2g04650</i>	ADP-glucose pyrophosphorylase family proteins	Seedling extremely small, not viable	0.72

The two genes (*At1g04300* and *At5g43560*) defective in the *DUPLO-2261* line code for tumor necrosis factor (TNF) receptor-associated factor (TRAF)-like proteins. Mammalian TRAF proteins have important signalling functions, especially as interacting partners of the TNF (Lee and Choi, 2007). While *DUPLO-2261* seedlings are only slightly smaller than WT seedlings, adult DMs are strongly stunted and bushier, although their siliques are WT-like in size and the plants are fertile (Table 2 and Figure 4d). Additionally, the efficiency of photosynthesis is reduced. Both genes seem to be expressed in all parts of the plant (*Arabidopsis* eFP Browser; Winter *et al.*, 2007), with *At1g04300* being expressed at higher levels in the later stages of seed development.

The *DUPLO-2268* line is defective for the two ubiquitin-specific proteases UBP12 (*At5g06600*) and 13 (*At3g11910*), which show very high similarity in both sequence (95.2% at protein level) and expression pattern (Pearson correlation coefficient (PCC) = 0.91). The DM plants are extremely small compared with the parental single mutants and produce few seeds, if any (Table 2 and Figure 4e). Plants have to be pre-grown on medium to allow survival on soil. Also the parental line *ubp12* is not able to grow on soil, whereas *ubp13* is only slightly smaller than WT. *DUPLO-2268/ubp12*

ubp13 resembles phenotypically the triple mutant *ubp15 ubp16 ubp17* (Liu *et al.*, 2008), which suggests additional redundancy within this gene family. Ewan *et al.* (2011) described *ubp12 ubp13* plants as being more susceptible to the virulent *Pseudomonas syringae* pv. tomato than WT, but displaying a WT-like growth behaviour. The loss-of-function allele for *UBP13* used by Ewan *et al.* (2011) differed from the one used in our study, a finding that might explain the discrepancies in the phenotypes observed.

The genes inactivated in *DUPLO-2589* encode eukaryote-specific aspartyl proteases (*At1g66180* and *At5g37540*). Genes for more than 50 A1-type aspartic proteases have been identified in *Arabidopsis* but the functions of their products are essentially unknown (Beers *et al.*, 2004). Some functions may be important for disease resistance and programmed cell death. The phenotype of the DM was not very pronounced, but the stature of the plants was slightly smaller than WT (Table 2 and Figure 4f).

The *DUPLO-2647* line corresponds to the DM *at1g51340 at3g08040*. *At3g08040* codes for the MATE efflux family protein MAN1 or FERRIC REDUCTASE DEFECTIVE3 (FRD3). FRD3 has previously been shown to control iron localization and to act as an efflux transporter of the efficient iron chelator citrate, and the *frd3-1* mutant has been characterised as

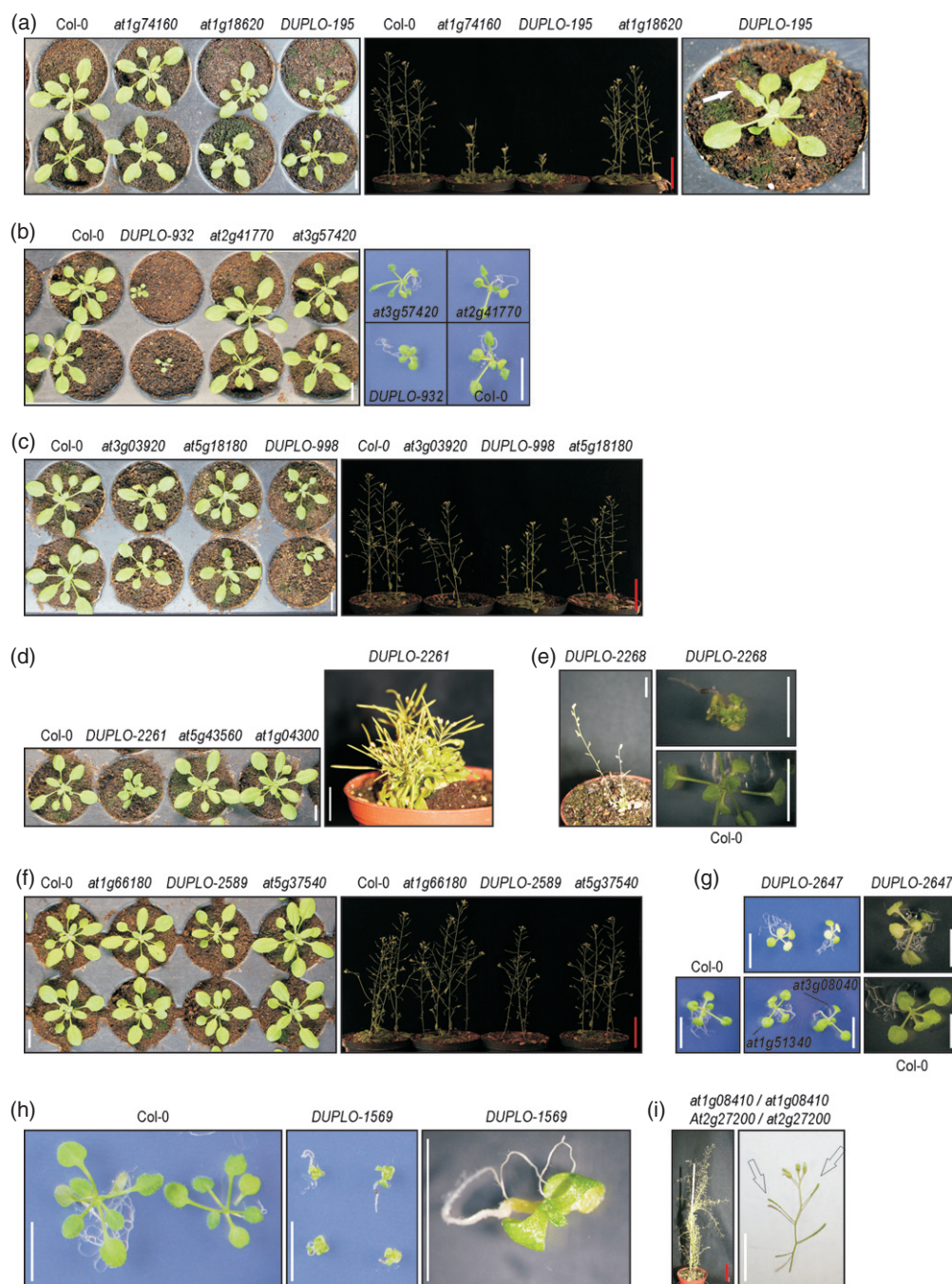


Figure 4. Phenotypes of selected DUPLO lines.

(a) *DUPLO-195* has a smaller rosette and shorter inflorescence compared with wild-type (WT) (Col-0). The phenotype is more severe than in the parental line *at1g74160*. Additionally, rolled leaves can be observed in the double knock-out mutants (DM) (see arrow).

(b) For *DUPLO-932* a smaller and more compact rosette can be observed.

(c) *DUPLO-998* is slightly smaller compared with the WT and the parental single-mutant lines.

(d) *DUPLO-2261* is only slightly smaller compared with WT at 30 dag (left panel), but the adult plant displays a stunted phenotype with fertile siliques (right panel).

(e) *DUPLO-2268* is dwarfed but capable of flowering and setting seeds if pre-grown on medium.

(f) *DUPLO-2589* is slightly smaller compared with WT and the parental single-mutant lines.

(g) A chlorotic phenotype can be observed for *DUPLO-2647*, which impairs autotrophic growth.

(h) *DUPLO-1569* displays strong growth retardation, which leads to seedling lethality.

(i) *DUPLO-810* plants, which are heterozygous for the *at2g27200* mutant allele, have many floral stems, but are infertile because the siliques do not develop (arrows).

Phenotypes were imaged on soil at the rosette stage (30 dag), at flowering (39 dag) or on media plates after 14 days. White bar = 1 cm, red bar = 5 cm.

Table 2 Physiological characterisation of mutant lines with phenotypes at the adult stage. At least six plants of DM lines, parental lines and WT were grown side-by-side

	Φ_{II}	Length of longest leaf (in % of WT)	Flowering time (dag)	Length of inflorescence (in % of WT at 39 dag)	Length of inflorescence (% of mature WT plants)
WT (Col-0)	0.72 ± 0.02	100.00 ± 19.74	28.0 ± 1.4	100.00 ± 10.1	100.0 ± 6.5
DUPLO-195	0.68 ± 0.03*	68.87 ± 8.96*	>31*	17.5 ± 12.5*	47.3 ± 17.6*
at1g18620	0.70 ± 0.02	97.49 ± 5.93	27.1 ± 0.5	103.1 ± 15.6	101.2 ± 7.5
at1g74160	0.72 ± 0.04	80.75 ± 6.08	>31*	17.1 ± 15.2*	49.9 ± 14.2*
DUPLO-932	0.71 ± 0.01	29.90 ± 10.4*	>31*	13.7 ± 4.6*	42.4 ± 5.9*
at3g57420	0.72 ± 0.02	99.04 ± 14.75	27.7 ± 2.2	121.4 ± 12.5	99.3 ± 4.8
at2g41770	0.74 ± 0.02	92.75 ± 11.69	27.2 ± 1.1	99.3 ± 15.8	91.3 ± 8.1
DUPLO-998	0.73 ± 0.01	73.94 ± 12.18*	>31*	85.6 ± 9.9*	98.2 ± 9.3
at5g18180	0.74 ± 0.02	83.77 ± 2.66	28.1 ± 0.4	104.1 ± 4.9	98.7 ± 6.3
at3g03920	0.70 ± 0.03	77.38 ± 23.11	26.6 ± 1.2	104.1 ± 14.8	86.0 ± 11.2
DUPLO-2261	0.69 ± 0.01*	75.29 ± 17.06*	27.2 ± 0.3	24.0 ± 9.4*	30.7 ± 4.0*
at5g43560	0.72 ± 0.01	97.79 ± 7.62	25.9 ± 1.3*	106.5 ± 18.1	89.9 ± 6.1
at1g04300	0.71 ± 0.03	100.32 ± 4.95	25.4 ± 1.4*	127.1 ± 16.8	93.1 ± 12.5
DUPLO-2268	nd	nd	nd	nd	nd
at5g06600	nd	nd	nd	nd	nd
at3g11910	0.72 ± 0.02	71.43 ± 17.15*	28.8 ± 2.6	77.3 ± 11.9*	84.4 ± 9.9*
DUPLO-2589	0.71 ± 0.01	100.53 ± 4.96	26.8 ± 1.3	99.0 ± 10.2	85.7 ± 12.1*
at5g37540	0.72 ± 0.03	105.92 ± 6.90	26.3 ± 2.1	95.1 ± 13.8	92.2 ± 16.2
at1g66180	0.74 ± 0.02	83.88 ± 7.63	>31*	80.7 ± 10.4*	94.4 ± 14.7
DUPLO-2647	nd	nd	nd	nd	nd
at1g51340	0.73 ± 0.01	98.03 ± 6.14	>31*	102.6 ± 12.6	100.2 ± 6.1
at3g08040	0.64 ± 0.06*	31.02 ± 7.79*	>31*	5.3 ± 10.0*	27.2 ± 18.1*

nd, not determined (because homozygous plants were not able to germinate directly on soil).

The length of the longest leaf (blade plus petiole) and Φ_{II} (as a marker for photosynthetic activity) were determined at 30 dag. Flowering time was measured in dag upon appearance of the bolt. The length of the inflorescence was measured after 39 dag and upon maturation.

Statistical significant differences from WT ($P < 0.05$) are indicated (*).

chlorotic and partially sterile (Green and Rogers, 2004; Roschttardt *et al.*, 2011). Our *frd3* mutant was also much smaller than WT, but showed normal fertility, although also the photosynthetic efficiency was reduced (Table 2). The T-DNA insertion in the DUPLO allele was close to the 3'-end of the *FRD3* gene, whereas it was more in the middle of the gene in the *frd3-1* allele. Both parental lines were flowering later than WT. *DUPLO-2647* was small and yellowish when grown on medium (Figure 4g), but some plants survived and set seeds after being transferred onto soil. This finding suggests that loss of the second homologous gene (*At1g51340*) has an additive effect on the mutant phenotype caused by the loss of *At3g08040*.

DUPLO lines that display lethality at the seedling stage

The second class of phenotypes observed in DUPLO lines is characterised by their failure to reach the reproductive phase, either because the seedling dies soon after germination or it is extremely small and stops growing prior to the adult phase ('seedling lethal' phenotype in Table 1). Six lines fall into this category and are described in the following paragraphs.

DUPLO-211 combines mutations in the *CVP1* (*At1g20330*) and *SMT3* (*At1g76090*) genes, which code for sterol 24-carbon methyltransferases that catalyse the synthesis of

structural sterols. The *cvp1 smt3* DM has been described before as being smaller and bushy, with short siliques and poor seed yield (Carland *et al.*, 2010). The more dramatic phenotype observed here might be associated with the fact that we used different mutant alleles.

In *DUPLO-713*, two genes (*HMGR1/At1g76490* and *HMGR2/At2g17370*) for 3-hydroxy-3-methylglutaryl CoA reductase (HMGR), which catalyses the first committed step in the mevalonate (MVA) pathway for isoprenoid biosynthesis in the cytoplasm, are inactivated. The *hmgr1* mutant was smaller than WT, but not as drastically dwarfed as reported previously (Suzuki *et al.*, 2004). DM plants germinated and grew into very small and inviable seedlings, whereas the DM described by Suzuki *et al.* (2009) showed male gametophyte lethality.

In the *DUPLO-1569* line, a gene pair coding for two tryptophan RNA-binding attenuator protein-like proteins (*At5g47420* and *At4g17420*), predicted to be localized in the chloroplast and containing at least three transmembrane domains, are inactivated. DM plants are very small and die early (Figure 4h). Inactivation of *At4g17420* alone in the heterozygous *at5g47420* background already causes a drastic reduction in plant size, and these very small plants do not flower.

In the *DUPLO-2098* line, genes for two proteins of unknown function (*At4g24330* and *At5g49945*) are disrupted. This line produces extremely small and inviable seedlings. The WT proteins are predicted to possess two transmembrane domains and to participate in the secretory pathway.

DUPLO-2490 (*at1g23310 at1g70580*) seedlings also die early when grown heterotrophically. *At1g23310* and *At1g70580* are both predicted to be localized in the peroxisome, and a glyoxylate aminotransferase activity has been assigned to each seedling.

In *DUPLO-2543*, genes for two ADP-glucose pyrophosphorylase family proteins (*At1g74910* and *At2g04650*) are disrupted. DM seedlings are extremely tiny and die after a few weeks even when grown on sucrose-containing media.

DUPLO lines that display lethality prior to germination

The third and largest group of lines are those that are unable to generate doubly mutant homozygotes. Even

re-segregation of at least 100 seeds on sucrose-containing medium led only to lines that remained heterozygous for one insertion and homozygous for the other. One must therefore assume that the DM dies at a very early stage, and this 'absence of homozygotes' suggests that the DM is either gametophytic or embryo lethal. For some of these DM lines (*DUPLO-151*, *-238*, *-1241*, *-2212* and *-2380*) lethality had been reported in the literature but using different alleles (Table 3). When the parental single mutants of DM lines that were incapable of producing homozygous DM embryos were re-investigated, it was found that in four cases (*DUPLO-110: at2g42910; DUPLO-2394: at2g24500; DUPLO-2666: at5g48950; DUPLO-2753: at4g08430*) already parental single mutants might be lethal.

About half of these gene pairs with essential functions seems to be important for metabolic processes. The proteins they encode are predicted to be phosphoribosyltransferases, UDP-galactose transporters, glycosyl hydrolases, nucleoside triphosphate hydrolases, sphingolipid

Table 3 List of DUPLO lines that are lethal before germination

Pair ID	Gene pair	Protein description	Phenotype	PCC
110	<i>At2g42910; At1g10700</i>	Phosphoribosyltransferase family protein; PRS3	Parent <i>At2g42910</i> may be lethal	0.69
151	<i>At2g02810; At1g14360</i>	UTR1; UTR3	Reyes <i>et al.</i> , 2010 (gametophytic lethal)	0.89
191	<i>At5g15870; At1g18310</i>	Glycosyl hydrolase family 81 proteins	–	0.37
213	<i>At1g76300; At1g20580</i>	SMD3; snRNP family protein	–	0.82
222	<i>At1g76850; At1g21170</i>	Exocyst complex components SEC5	–	0.74
238	<i>At4g10040; At1g22840</i>	Cytochrome C-2; Cytochrome C-1	Welchen <i>et al.</i> , 2012 (arrest of embryo development)	0.74
601	<i>At1g21190; At1g76860</i>	snRNP family proteins	–	0.84
810	<i>At1g08410; At2g27200</i>	P-loop containing nucleoside triphosphate hydrolases superfamily proteins	–	0.70
1022	<i>At1g16180; At3g06170</i>	Serine-domain containing serine and sphingolipid biosynthesis proteins	–	0.85
1241	<i>At5g40650; At3g27380</i>	Succinate dehydrogenase 2-2; Succinate dehydrogenase 2-1	León <i>et al.</i> , 2007 (essential for gametophyte development)	0.92
1361	<i>At2g38700; At3g54250</i>	MVD1; GHMP kinase family protein	–	0.86
1545	<i>At3g22440; At4g14900</i>	FRIGIDA-like proteins	–	0.87
1641	<i>At5g55120; At4g26850</i>	VTC5; VTC2	Dowdle <i>et al.</i> , 2007 (seedling lethal)	0.66
2212	<i>At5g09740; At5g64610</i>	HAM2; HAM1	Latrasse <i>et al.</i> , 2008 (lethal)	0.86
2253	<i>At1g53850; At3g14290</i>	PAE1; PAE2	–	0.95
2257	<i>At5g66140; At3g51260</i>	PAD2; PAD1	–	0.94
2341	<i>At1g17890; At1g73250</i>	GER2; GER1	–	0.85
2349	<i>At1g74040; At1g18500</i>	IPMS2; IPMS1	–	0.85
2361	<i>At5g60550; At3g45240</i>	GRIK2; GRIK1	–	0.84
2380	<i>At1g65660; At4g37120</i>	SMP1; SMP2	Clay and Nelson, 2005 (lethal)	0.83
2394	<i>At2g24500; At4g31420</i>	FZF; Zinc-finger protein 622	Parent <i>at2g24500</i> may be lethal	0.82
2666	<i>At1g48320; At5g48950</i>	Thioesterase superfamily proteins	Parent <i>at5g48950</i> may be lethal	0.59
2753	<i>At5g45570; At4g08430</i>	Ulp1 protease family proteins	Parent <i>at4g08430</i> may be lethal	0.49

PCC, Pearson correlation coefficient.

For all these lines, no homozygous double knock-out mutants (DM) plant could be obtained. Additional information, including the used T-DNA insertion alleles are provided in Table S1.

biosynthesis proteins, succinate dehydrogenases, mevalonate diphosphate decarboxylases, guanylyltransferases, histone acetyltransferases, NAD(P)-binding Rossmann-fold superfamily proteins, 2-isopropylmalate synthases and thioesterases (Table 3). Other gene products are constituents of multiprotein complexes such as cytochrome c, snRNP core proteins, exocyst complex components, small nuclear ribonucleoprotein family proteins, subunits of proteasomes and pre-mRNA splicing interacting factor. Only a few proteins, such as FRIGIDA-like proteins, kinases, Ulp1 protease family proteins and zinc-finger proteins, seem to be involved in signal transduction.

In *DUPLO-810*, genes *At2g27200* and *At1g08410* for two members of the P-loop-containing nucleoside triphosphate hydrolase superfamily are disrupted. Each of the parental single mutants was partially sterile (Figure 4i).

DUPLO lines that show phenotypes under non-standard conditions or at the molecular level

DUPLO lines that do not exhibit a visually discernible phenotype under greenhouse conditions might still display a

phenotype when exposed to special conditions (such as stress or special lighting conditions) or at the molecular level (for instance altered transcript or metabolite profiles). The prevalence of such lines in the DUPLO collection was assessed on the basis of literature searches. These lines revealed that, for 14 DMs with counterparts in the DUPLO collection, phenotypes had been identified which would not have been detectable in our initial phenotypic screen of greenhouse-grown plants (Table 4). This finding indicates that the detailed analysis of DMs can lead to the discovery of additional phenotypes. For example, the DM *pif4 pif5* (*DUPLO-2409*) shows a hypersensitive phenotype at lower fluence rates of far-red light (Lorrain *et al.*, 2009), whereas the DM *far1 fhy3* (*DUPLO-1196*) has an elongated hypocotyl under far-red light, even longer than those hypocotyls seen in each of the single-mutant parents (Wang and Deng, 2002). The counterpart of the *DUPLO-1202* line showed no visible phenotype under our growth conditions but, in a more detailed analysis, Rossini *et al.* (2006) have demonstrated that the DM shows a slight reduction in the chlorophyll content in mature leaves and a less

Table 4 List of DUPLO lines displaying a phenotype under non-standard conditions or at the molecular level

Pair ID	Gene pair	Protein description	Phenotype	References	PCC
148	<i>At2g02950</i> ; <i>At1g14280</i>	PKS1; PKS2	Involved in phyA-mediated VLFR	Lariguet <i>et al.</i> (2003)	0.74
380	<i>At3g19280</i> ; <i>At1g49710</i>	FucTA; FucTB	Variation of the complex-type N-glycans	Strasser <i>et al.</i> (2004)	0.68
1196	<i>At4g15090</i> ; <i>At3g22170</i>	FAR1; FHY3	Elongated hypocotyl under far-red light	Wang and Deng (2002)	0.85
1202	<i>At4g14690</i> ; <i>At3g22840</i>	ELIP2; ELIP1	Reduced chlorophyll content, less accumulation of zeaxanthin in high light; altered germination under abiotic stress	Rossini <i>et al.</i> (2006), Rizza <i>et al.</i> (2011)	0.79
1210	<i>At2g48110</i> ; <i>At3g23590</i>	REF4; RFR1	Enhanced expression of phenylpropanoid biosynthetic genes, increased accumulation of downstream products	Bonawitz <i>et al.</i> (2012)	0.75
2376	<i>At1g51590</i> ; <i>At3g21160</i>	MNS1; MNS2	Aberrant Man(8)GlcNAc(2) accumulation	Liebmingner <i>et al.</i> (2009)	0.83
2409	<i>At2g43010</i> ; <i>At3g59060</i>	PIF4; PIL6	Hypersensitive to lower fluence rates of far-red light	Lorrain <i>et al.</i> (2009)	0.81
2412	<i>At4g31920</i> ; <i>At2g25180</i>	ARR10; ARR12	Inhibition of root elongation, green callus formation from root explants	Yokoyama <i>et al.</i> (2007)	0.81
2429	<i>At5g45110</i> ; <i>At4g19660</i>	NPR3; NPR4	Elevated PR-1 expression, enhanced resistance to virulent bacterial and oomycete pathogens	Zhang <i>et al.</i> (2006)	0.80
2440	<i>At1g50890</i> ; <i>At4g27060</i>	SP2L; CN, SPR2, TOR1	Enhanced right-handed helical growth	Yao <i>et al.</i> (2008)	0.79
2474	<i>At5g47230</i> ; <i>At4g17490</i>	ERF5; ERF6	Increased susceptibility to <i>B. cinerea</i> , JA-induced gene expression reduced	Moffat <i>et al.</i> (2012)	0.77
2488	<i>At2g31650</i> ; <i>At1g05830</i>	ATX1; ATX2	Involved in <i>FLC</i> activation and histone methylation	Yun <i>et al.</i> (2012)	0.76
2653	<i>At1g27320</i> ; <i>At2g01830</i>	AHK3; CRE1	Reduced responsiveness to cytokinin	Franco-Zorrilla <i>et al.</i> (2005), Riefler <i>et al.</i> (2006)	0.60
2668	<i>At1g77130</i> ; <i>At3g18660</i>	GUX3; GUX1	Shorter GlcA and MeGlcA side chains on xylan in the cell wall	Lee <i>et al.</i> (2012)	0.59

PCC, Pearson correlation coefficient.

accumulation of zeaxanthin in high light conditions. For the DM corresponding to *DUPLO-380* structural variations in complex-type *N*-glycans have been detected using MALDI-ToF mass spectrometry (Strasser *et al.*, 2004).

The relationship between mutant phenotype frequency and gene pair characteristics

Taking all phenotypes together, we were able to attribute some phenotypic and/or developmental deviation from the WT to 50 of the 200 DUPLO lines investigated. For all DUPLO gene pairs, a multidimensional Arabidopsis mRNA expression dataset comprising 1765 expression arrays was used to rank these pairs according to the degree of similarity between their expression profiles by global correlation. The expression dataset covers a large range of biological processes, including hormone and light treatments, mutant analysis, biotic and abiotic stresses, various tissues and developmental stages (Haberer *et al.*, 2006).

To investigate whether a correlation exists between similarity of expression pattern and the frequency of DM mutant phenotypes observed, the frequency of DM phenotypes was plotted against the PCC values for the corresponding gene pairs (Figure 5). This comparison clearly showed that gene pairs with very similar expression profiles, i.e. a high PCC value, are more likely to result in DM phenotypes than gene pairs with lower PCC values. Similarly, gene pairs with high sequence similarity also displayed a trend towards a high mutant phenotype frequency (Figure 5). When we correlated expression similarity and coding sequence similarity of the gene pairs, we obtained PCC values of 0.40 (for the 50 DUPLO lines with phenotypes) and 0.24 (for all 200 DUPLO lines), indicating that high expression similarity is not necessarily associated with high coding sequence similarity and vice versa. In consequence, our data suggest that duplicated genes that display high coding sequence similarity or expression profiles more frequently result in DM phenotypes.

Public availability of DUPLO lines

The DUPLO lines are available through NASC, together with the corresponding parental single-mutant alleles. For DUPLO lines that are not viable or do not produce sufficient seed (see above and Table S1), the segregating F2 or F3 population will be distributed. Information on the parental alleles, primers used for genotyping and the initial phenotypic evaluation will be available at <http://www.gabi-kat.de/duplo.html> (see Figure S3).

DISCUSSION

Gene knock-outs are an important tool in the study of gene functions in the model plant *A. thaliana* (reviewed in Bolle *et al.*, 2011), but when additional gene copies with overlapping or even redundant functions exist their utility is limited. About one-sixth of the genes in the Arabidopsis genome are present in multiple copies which may have overlapping or even fully redundant functions. In such cases, single-gene knock-outs are likely to be uninformative. Because we selected pairs of clearly paralogous genes that have no other close relatives in the genome, the function(s) they mediate should be fully inactivated in the DUPLO lines, as interference due to functional overlap with a third gene is quite unlikely. Practical applications of DUPLO lines include the phenotypic characterisation of individual DMs in reverse genetics approaches, but also extend to systematic forward genetics screens such as those already established for conventional (single) mutant collections. Because most of the unwanted secondary mutations have been segregated out in the F2 and F3 generations, phenotypic effects unrelated to the inactivated gene pairs should be rare.

Comparison with published data

When the phenotypes observed in this study were compared with published results, some differences were

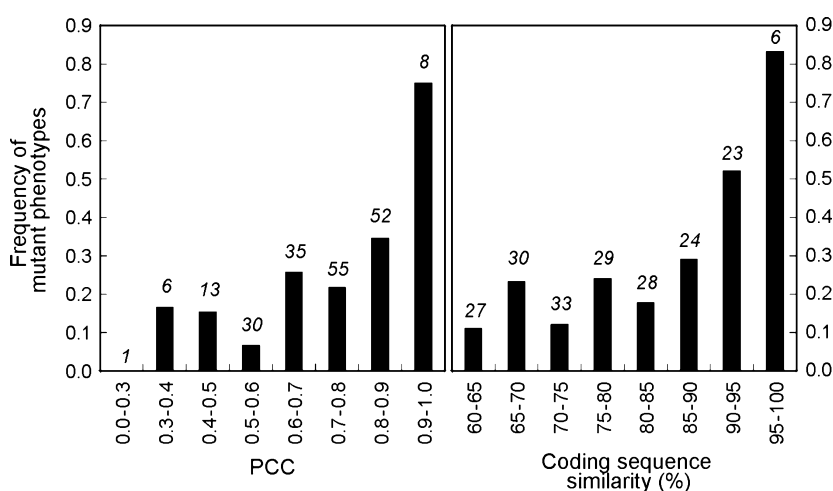


Figure 5. Effects of coding sequence or expression pattern similarity of gene pairs on double knock-out mutants (DM) phenotype frequency. In the left (right) panel, the frequency of DM phenotypes among the 200 DUPLO lines is plotted against the Pearson correlation coefficient (PCC) value (sequence similarity) for the members of each gene pair. The number of lines assigned to the particular intervals of PCC values and coding sequence similarities is given above each column.

observed. In several cases we observed less severe mutant phenotypes than those described previously. Thus, the lines *DUPLO-2332* and *-2572* and their parental single mutants were viable, despite the fact that embryo lethality had previously been reported for other alleles of one parental single mutant of each gene pair (*DUPLO-2332*: *SCO3*, Albrecht *et al.*, 2010; *DUPLO-2572*: *BUB3.1*, Lermontova *et al.*, 2008). Similarly, for one of the two genes mutated in *DUPLO-2693* (*At1g76620*), which codes for a protein of unknown function, a mutant phenotype with altered seed pigmentation was observed (Bryant *et al.*, 2011; *pde339*). In our hands, *DUPLO-2693* was WT-like.

Conversely, in several cases we observed more severe mutant phenotypes than reported before. Thus the phenotype of *DUPLO-2268* (very small and barely viable on soil) was more severe than the one described by Ewan *et al.* (2011), who found no obvious growth defect. This difference might be related to the fact that different alleles were used for *UBP12* and *UBP13*, although the DUPLO insertions were closer to the 3' end of the coding regions than the insertions in the allele used by Ewan *et al.* (2011). Also for *DUPLO-211* (*cvp1 smt3*) we observed a more pronounced phenotype: here seedlings were lethal, whereas the DM described by Carland *et al.* (2010) (which carried a different *smt3* and *cvp1* allele; the *cvp1* allele was derived from EMS mutation screen in contrast to the DUPLO lines which carries a T-DNA insertion close to the 3'-end of the gene) was characterised by discontinuities in the vein pattern of the cotyledon, defective root growth, loss of apical dominance, sterility, and homeotic floral transformations. The *vtc2 vtc5* line, analysed by Dowdle *et al.* (2007), is disrupted at a step in the ascorbate biosynthetic pathway and shows growth arrest immediately upon germination and the cotyledons subsequently bleach. Its counterpart, *DUPLO-1641*, generated from two different alleles, was lethal before germination in our hands, and even supplementation with L-galactose failed to rescue seedlings. In this case the insertion in our allele used for *vtc5* was closer to the ATG codon compared with the ones used by Dowdle *et al.* (2007), although localized within an intron. The allele used for *vtc2* by Dowdle *et al.* (2007) had a single base exchange at the predicted 3' splice site of intron 5, which produces a truncated mRNA at a reduced level, whereas the DUPLO allele contains a T-DNA insertion between the sixth and seventh exon. The alleles used to generate *DUPLO-713* also differ from those examined by Suzuki *et al.* (2009): for *hmg1* both insertions are within the first exon, but for *hmg2* the allele used by Suzuki *et al.* (2009), was in the third exon, whereas the DUPLO allele had an insertion in the first exon. This finding could explain the difference between the phenotypes found – gametophytic lethality in the earlier report and seedling lethality in this study.

In summary, we found eight instances in which the DM phenotypes observed in this study differed from those

described before. The most plausible explanation for such differences is that they result from the use of different mutant alleles as insertions within the intron or close to the 3'-part of the coding region might not lead to complete knock-out lines. This situation in turn implies that, for DMs also, combinations of at least two different single-mutant alleles have to be analysed before one can unambiguously assign functions to gene pairs. In cases of discrepancies like those reported above, a third independent DM line needs to be analysed, together with all available single and DMs, under the same growth conditions.

Frequency of mutant phenotypes

Our initial phenotyping analysis showed that 13 (or 6.5%) of the 200 DUPLO lines display a visually discernible phenotype (adult phenotype and seedling lethal) under greenhouse conditions. This is higher than the frequency of visible phenotypes observed among 4000 lines with transposon insertions in their coding regions; there 140 lines (or 3.5%) with visually discernible phenotypes were found (Kurumori *et al.*, 2006). Although the number of DUPLO lines is still too small to allow robust statistics, this difference (by a factor of almost two) might indicate that the DUPLO collection is enriched for lines with phenotypic variations (adult phenotypes and seedling lethal) when compared with single-gene mutant collections. Moreover, we observed 23 lines that were lethal before germination and 14 lines for which phenotypes that become manifest under non-standard conditions or at the molecular level had been reported in the literature. Thus, 50 DUPLO lines (or 25%) displayed a mutant phenotype. Because the distribution of gene functions in the set of 2108 gene pairs (from which the 200 DUPLO lines were selected) was similar to that for the whole genome (see Figure 2), we speculate that these 200 gene pairs were enriched for housekeeping functions which, when mutated, result in obvious phenotypes visible in our initial screen (see below). Within the set of 50 lines with phenotypic deviations from WT, lines that contained mutated gene pairs with either highly similar expression pattern or sequence were clearly overrepresented (see Figure 5). Because we found no strong correlation between high coding sequence similarity and highly similar expression profiles, it appears that promoter and coding regions of gene pairs can display different degrees of diversification. Assuming that the 50 gene pairs with clear mutant phenotypes are enriched for housekeeping functions and possess either similar promoter regions (resulting in similar expression profiles) or similar coding sequences, we speculate that housekeeping genes are under a special type of post-duplication selection. This type of selection conserves promoter or coding regions, or sometimes both (see PCC of 0.40 for sequence similarity and expression similarity in case of the 50 gene pairs resulting in phenotypes), and might serve to stabilise

the expression level of gene products with housekeeping functions.

EXPERIMENTAL PROCEDURES

Selection of candidate genes

To identify paralogous genes coding for very similar proteins, we conducted an all-against-all BlastP comparison of all *A. thaliana* genes (TAIR version 6) and selected for gene pairs with an e-value $\leq 10^{-10}$. Preselected pairs were subjected to an optimal global alignment and paralogues with <20% gaps and a minimal alignment similarity $\geq 60\%$ were retained. These relations were projected onto a similarity graph, on which genes were represented by nodes and retained gene pairs were connected by edges. Paralogous gene clusters were computed as connected components. Gene pairs that were <7.5 Mbp apart in the genome were not considered further. This step removed tandemly repeated and genetically closely linked genes, for which it would be difficult to obtain homozygous DMs by genetic crosses. The availability either in the GABI-Kat or SALK collection of alleles with an insertion within the coding region was a prerequisite for the generation of DM lines in the Col-0 background. During the course of the project the annotation was updated to version 10 of TAIR, and the locations of insertions in lines assigned to DU-PLO pairs were re-evaluated. Insertions outside of the coding region or the 5'-UTR were swapped for better alleles, if available.

To evaluate the over- and under-representation of genes in the collection, the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 was used (Huang da *et al.*, 2009a, b). The graphs for functional categorisation were generated using the TAIR annotation tool (<http://www.arabidopsis.org/tools/bulk/go/index.jsp>).

Expression similarities were calculated as PCCs from an all-against-all comparison of 170 microarray experiments (1765 individual expression arrays) based on the Affymetrix ATH1 platform. The experiments comprise the AtGenExpress and the NASC release three datasets (<http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl>). Probe sets were realigned to the TAIR version 6 annotation and analysed as described previously (Haberer *et al.*, 2006).

Growth conditions

All lines were grown on standard soil in the greenhouse. Plants were fertilised with a liquid nitrogen-phosphate-potassium fertiliser as recommended by the manufacturer. For growth on agar medium, seeds were surface-sterilised and sown on agar plates containing 1× MS (Sigma-Aldrich, <http://www.sigmaaldrich.com/germany.html>), 3% sucrose, 0.01% myo-inositol, 0.05% MES, vitamins (biotin, nicotinic acid, pyridoxine, thiamine). For selection of T-DNA-containing T2 plants of GABI-Kat lines, 5.25 µg ml⁻¹ sulfadiazine was added to the plates. Segregation of the F2 or F3 generation was in part performed on 1× MS medium including 1% sucrose.

Selection of primers

For each individual T-DNA insertion allele, an allele-specific primer was designed using scripts based upon the program PRIMER3 (Rozen and Skaletsky, 2000), taking into account the predicted T-DNA insertion position, the orientation of the T-DNA insertion, a desired amplicon size of 400–760 bp, and a T_m of about 60.5°C. The allele-specific primers were used together with T-DNA border primers, which are specific either for the GABI-Kat or SALK

T-DNA. We refer to the resulting fragments as 'confirmation amplicons'. The T-DNA insertion positions were deduced from the sequences of the confirmation amplicons (Kleinboelting *et al.*, 2012). Amplicons designed to show presence/absence of the wild-type alleles ('genotyping amplicons') are based on primer pairs that span the deduced insertion sites. In this case, one of the allele-specific primers for the confirmation amplicon was routinely used together with a primer based on genomic sequences located at the opposite side of the insertion and designed with PRIMER3 (Figure S2). The size range of the genotyping amplicons was 540–1260 bp. If synthesis of the initially designed amplicon failed, primers were redesigned. All primers shown in this work and on the website were experimentally verified.

Genotyping

Genomic DNA of the parental lines was prepared from ground leaf material with a modified CTAB-DNA protocol (Dellaporta *et al.*, 1983; Jobes *et al.*, 1995). Genomic DNA of the following generations was extracted from 5 to 10 mg of leaf material from a 3- to 5-week-old plant. After addition of 400 µl of extraction buffer (0.2 M Tris/HCl pH 7.5; 0.25 M NaCl; 0.025 M EDTA; 0.5% SDS) and a stainless steel ball (5 mm diameter) the leaf material was homogenised by shaking the tubes for 3 min at 30 Hz in a Retsch homogeniser (Retsch Mixer Mill 300; Qiagen, <http://www.qiagen.com>). The suspension was centrifuged at 13 000 g and the supernatant was transferred into a new tube. An equal volume of isopropanol was added, mixed, and the mixture was centrifuged after a 2-min incubation at room temperature. The supernatant was transferred to a new tube and 2.5 volumes of 70% ethanol were added. After mixing and centrifugation the supernatant was removed and the pellet was dried for 30 min. Finally, the DNA was dissolved in 100 µl of H₂O by shaking at room temperature. DNA was stored at –20°C. For GABI-Kat lines the template DNA was tested for the presence of a T-DNA insertion with the SUL^r ORF-specific primers Sul2 and Sul4 (see Figure S2). For SALK lines, the SALK T-DNA-specific primers G171 and G172 were used (see Figure S2).

PCR was performed using either the primers for the confirmation or the genotyping amplicon. PCR was carried out with 1–2 µl of genomic DNA in a final reaction volume of 50 µl containing 17 mM MgCl₂. Col-0 DNA was used for control of the genotyping amplicon PCR (positive result expected) and confirmation amplicon PCR (negative result expected). For the confirmation of FST-based insertion-site predictions, genomic DNA isolated from a pool of up to 25 seedlings was used as template in the confirmation amplicon PCRs. For GABI-Kat lines, the border PCR primers for the confirmation amplicons were 8474 for left border (LB) FSTs and 3144 for right border (RB) FSTs. For SALK lines the LB primer was Lbb3. If a confirmation PCR failed, a second reaction was performed using a different gene-specific primer. Confirmation amplicons were purified with the ExoSAP-IT kit (USB, USA) and sequenced with a gene-specific primer and a border primer. The border sequencing primers were 8409 for LB FSTs and 3144 for RB FSTs in the case of GABI-Kat lines and R210 in the case of SALK lines (LB). The gene-specific primers were those used for the PCRs. Sequencing results were compared to the genomic sequence of *A. thaliana* (TAIR version 10 genome sequence and annotation dataset) using the BLAST algorithm (Altschul *et al.*, 1990). When the insertion-site prediction for at least one of the two sequences was consistent with the FST-based prediction, the respective allele was regarded as confirmed in the line. If no confirmation amplicon could be obtained, a suitable replacement was chosen, when further lines

were available. After confirmation of the insertions and the establishment of the genotyping amplicon using Col-0 DNA as described above, genomic DNA of 12–18 individual T2 plants was prepared and the plants were genotyped in order to identify predominantly homozygous parental lines.

Of the parental lines, four to eight plants were analysed with the insertion- and gene-specific primers prior to crossing, and predominantly homozygous parents were selected for crossing. Of the F1 generation 2–4 reciprocally crossed plants were tested for both insertions. For the F2 generation 54 plants were tested with the confirmation and genotyping amplicons of both insertions. DMs were confirmed in the F3 or F4 generation. Segregation on plates made with sucrose-containing and sulfadiazine-containing medium (for GABI-Kat lines) allowed for calculation of the segregation rate of germinated versus non-germinated seeds or non-viable seedlings. For statistical analysis about 100 seeds were used.

Phenotypic analysis

Phenotypic analysis was performed in the F3 or F4 generation. Three to six plants per DM lines were grown in parallel to the WT. The plants were inspected visually after 1, 3 and 5 weeks. Rosette size, leaf coloration, flowering time, overall height, branching pattern and floral development were noted. For confirmation of the observed phenotype at least six DM plants were grown side-by-side with the parental lines and WT. Three-week-old seedlings were photographed to measure the size of the rosette indicated by the length of the longest leaf (blade plus petiole) and the effective quantum yield of photosystem II (Φ_{II}) was measured as a marker for photosynthetic activity (Leister *et al.*, 1999). Flowering time was measured in days after germination (dag) upon appearance of the bolt. The overall height of the flowering stem was measured after 39 dag and upon maturation. Additional abnormalities were documented by photographs.

In-silico analyses

To compare the expression in different tissues and developmental stages for the individual gene pairs, the eFP Browser was used (Winter *et al.*, 2007). Prediction of subcellular localization and cluster analysis of related plant protein sequences were performed using the Aramemnon database (Schwacke *et al.*, 2003).

ACKNOWLEDGEMENTS

We thank Mario Rosso and Yong Li for their effort to start this project and the selection of the alleles, the BMBF for funding (grant GABI-DUPLO; Förderkennzeichen: 0315055A) and Paul Hardy for critical reading of the manuscript. We acknowledge the technical assistance of Elvira Glatt, Renate Harder, Christine Gonzales-Serrano, Simona Horn, Michael Mulatsch, Susanne Olbrich, Helene Schellenberg, Laura Schröder, Nina Schmidt, Andrea Voigt and Marja-Leena Wilke.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Functional distribution of proteins covered by the DUPLO collection.

Figure S2. Overview of the primers used to analyse T-DNA insertion lines.

Figure S3. Screenshots of the DUPLOdb.

Table S1. List of the 200 DUPLO lines generated and distributed to NASC so far (as of March 2013).

REFERENCES

- Abel, S., Savchenko, T. and Levy, M. (2005) Genome-wide comparative analysis of the IQD gene families in *Arabidopsis thaliana* and *Oryza sativa*. *BMC Evol. Biol.* **5**, 72.
- AGI. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
- Albrecht, V., Simková, K., Carrie, C., Delannoy, E., Giraud, E., Whelan, J., Small, I.D., Apel, K., Badger, M.R. and Pogson, B.J. (2010) The cytoskeleton and the peroxisomal-targeted snowy cotyledon3 protein are required for chloroplast development in *Arabidopsis*. *Plant Cell*, **22**, 3423–3438.
- Alonso, J.M., Stepanova, A.N., Leisse, T.J. *et al.* (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, **301**, 653–657.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
- Armisen, D., Lecharny, A. and Aubourg, S. (2008) Unique genes in plants: specificities and conserved features throughout evolution. *BMC Evol. Biol.* **8**, 280.
- Beers, E.P., Jones, A.M. and Dickerman, A.W. (2004) The S8 serine, C1A cysteine and A1 aspartic protease families in *Arabidopsis*. *Phytochemistry*, **65**, 43–58.
- Blanc, G., Hokamp, K. and Wolfe, K.H. (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* **13**, 137–144.
- Bolle, C., Schneider, A. and Leister, D. (2011) Perspectives on Systematic Analyses of Gene Function in *Arabidopsis thaliana*: New Tools, Topics and Trends. *Curr. Genomics*, **12**, 1–14.
- Bonawitz, N.D., Soltan, W.L., Blatchley, M.R., Powers, B.L., Hurlock, A.K., Seals, L.A., Weng, J.K., Stout, J. and Chapple, C. (2012) REF4 and RFR1, subunits of the transcriptional coregulatory complex mediator, are required for phenylpropanoid homeostasis in *Arabidopsis*. *J. Biol. Chem.* **287**, 5434–5445.
- Bowers, J.E., Chapman, B.A., Rong, J. and Paterson, A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, **422**, 433–438.
- Bryant, N., Lloyd, J., Sweeney, C., Myouga, F. and Meinke, D. (2011) Identification of nuclear genes encoding chloroplast-localized proteins required for embryo development in *Arabidopsis*. *Plant Physiol.* **155**, 1678–1689.
- Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D. and May, G. (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* **4**, 10.
- Carland, F., Fujioka, S. and Nelson, T. (2010) The sterol methyltransferases SMT1, SMT2, and SMT3 influence *Arabidopsis* development through nonbrassinosteroid products. *Plant Physiol.* **153**, 741–756.
- Charon, C., Bruggeman, Q., Thareau, V. and Henry, Y. (2012) Gene duplication within the Green Lineage: the case of TEL genes. *J. Exp. Bot.* **63**, 5061–5077.
- Clay, N.K. and Nelson, T. (2005) The recessive epigenetic swellmap mutation affects the expression of two step II splicing factors required for the transcription of the cell proliferation gene STRUWELPETER and for the timing of cell cycle arrest in the *Arabidopsis* leaf. *Plant Cell*, **17**, 1994–2008.
- DalCorso, G., Pesaresi, P., Masiero, S., Aseeva, E., Schunemann, D., Finazzi, G., Joliot, P., Barbato, R. and Leister, D. (2008) A complex containing PGR1 and PGR5 is involved in the switch between linear and cyclic electron flow in *Arabidopsis*. *Cell*, **132**, 273–285.
- Dellaporta, S., Wood, J. and Hicks, J.B. (1983) A plant DNA miniprep. *Plant Mol. Biol. Rep.* **1**, 19–21.
- Dowdle, J., Ishikawa, T., Gatzek, S., Rolinski, S. and Smirnov, N. (2007) Two genes in *Arabidopsis thaliana* encoding GDP-L-galactose phosphorylase are required for ascorbate biosynthesis and seedling viability. *Plant J.* **52**, 673–689.
- Drevensk, S., Goussot, M., Duroc, Y. *et al.* (2012) The *Arabidopsis* TRM1-TON1 interaction reveals a recruitment network common to plant cortical microtubule arrays and eukaryotic centrosomes. *Plant Cell*, **24**, 178–191.
- Ewan, R., Pangestuti, R., Thornber, S., Craig, A., Carr, C., O'Donnell, L., Zhang, C. and Sadanandom, A. (2011) Deubiquitinating enzymes AtUBP12 and AtUBP13 and their tobacco homologue NtUBP12 are negative regulators of plant immunity. *New Phytol.* **191**, 92–106.

- Franco-Zorrilla, J.M., Martin, A.C., Leyva, A. and Paz-Ares, J. (2005) Interaction between phosphate-starvation, sugar, and cytokinin signaling in Arabidopsis and the roles of cytokinin receptors CRE1/AHK4 and AHK3. *Plant Physiol.* **138**, 847–857.
- Green, L.S. and Rogers, E.E. (2004) FRD3 controls iron localization in Arabidopsis. *Plant Physiol.* **136**, 2523–2531.
- Gupta, R., He, Z. and Luan, S. (2002) Functional relationship of cytochrome c_6 and plastocyanin in Arabidopsis. *Nature*, **417**, 567–571.
- Haberer, G., Mader, M.T., Kosarev, P., Spannagl, M., Yang, L. and Mayer, K.F. (2006) Large-scale cis-element detection by analysis of correlated expression and sequence conservation between Arabidopsis and *Brassica oleracea*. *Plant Physiol.* **142**, 1589–1602.
- He, X.L. and Zhang, J.Z. (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, **169**, 1157–1164.
- Huang da, W., Sherman, B.T., Zheng, X., Yang, J., Imamichi, T., Stephens, R. and Lempicki, R.A. (2009a) Extracting biological meaning from large gene lists with DAVID. *Curr. Protoc. Bioinformatics*, Chapter 13, Unit 13.11.
- Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57.
- Ihnatowicz, A., Pesaresi, P., Varotto, C., Richly, E., Schneider, A., Jahns, P., Salamini, F. and Leister, D. (2004) Mutants for photosystem I subunit D of *Arabidopsis thaliana*: effects on photosynthesis, photosystem I stability and expression of nuclear genes for chloroplast function. *Plant J.* **37**, 839–852.
- Ihnatowicz, A., Pesaresi, P. and Leister, D. (2007) The E subunit of photosystem I is not essential for linear electron flow and photoautotrophic growth in *Arabidopsis thaliana*. *Planta*, **226**, 889–895.
- Jiang, D., Yin, C., Yu, A., Zhou, X., Liang, W., Yuan, Z., Xu, Y., Yu, Q., Wen, T. and Zhang, D. (2006) Duplication and expression analysis of multicopy miRNA gene family members in Arabidopsis and rice. *Cell Res.* **16**, 507–518.
- Jiao, Y., Wickett, N.J., Ayyampalayam, S. et al. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 97–100.
- Jobs, D.V., Hurley, D.L. and Thien, L.B. (1995) Plant DNA isolation: a method to efficiently remove polyphenolics, polysaccharides, and RNA. *Taxon*, **44**, 379–386.
- Kleinboelting, N., Huep, G., Kloetgen, A., Viehoveer, P. and Weissshaar, B. (2012) GABI-Kat SimpleSearch: new features of the *Arabidopsis thaliana* T-DNA mutant database. *Nucleic Acids Res.* **40**, D1211–D1215.
- Kolkusaoglu, H.U., Bovet, L., Klein, M., Eggmann, T., Geisler, M., Wanke, D., Martinoia, E. and Schulz, B. (2002) Family business: the multidrug-resistance related protein (MRP) ABC transporter genes in *Arabidopsis thaliana*. *Planta*, **216**, 107–119.
- Kong, H., Landherr, L.L., Frohlich, M.W., Leebens-Mack, J., Ma, H. and dePamphilis, C.W. (2007) Patterns of gene duplication in the plant SKP1 gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth. *Plant J.* **50**, 873–885.
- Kuromori, T., Wada, T., Kamiya, A. et al. (2006) A trial of phenome analysis using 4000 Ds-insertional mutants in gene-coding regions of Arabidopsis. *Plant J.* **47**, 640–651.
- Lariguet, P., Boccalandro, H.E., Alonso, J.M., Ecker, J.R., Chory, J., Casal, J.J. and Fankhauser, C. (2003) A growth regulatory loop that provides homeostasis to phytochrome a signaling. *Plant Cell*, **15**, 2966–2978.
- Latrasse, D., Benhamed, M., Henry, Y., Domenichini, S., Kim, W., Zhou, D.X. and Delarue, M. (2008) The MYST histone acetyltransferases are essential for gametophyte development in Arabidopsis. *BMC Plant Biol.* **8**, 121.
- Lee, S.Y. and Choi, Y. (2007) TRAF1 and its biological functions. *Adv. Exp. Med. Biol.* **597**, 25–31.
- Lee, C., Teng, Q., Zhong, R. and Ye, Z.H. (2012) Arabidopsis GUX proteins are glucuronyltransferases responsible for the addition of glucuronic acid side chains onto xylan. *Plant Cell Physiol.* **53**, 1204–1216.
- Leister, D. (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. *Trends Genet.* **20**, 116–122.
- Leister, D., Varotto, C., Pesaresi, P., Niwergall, A. and Salamini, F. (1999) Large-scale evaluation of plant growth in *Arabidopsis thaliana* by non-invasive image analysis. *Plant Physiol. Biochem.* **37**, 671–678.
- León, G., Holuigue, L. and Jordana, X. (2007) Mitochondrial complex II is essential for gametophyte development in Arabidopsis. *Plant Physiol.* **143**, 1534–1546.
- Lermontova, I., Fuchs, J. and Schubert, I. (2008) The Arabidopsis checkpoint protein Bub3.1 is essential for gametophyte development. *Front Biosci.* **13**, 5202–5211.
- Liebminger, E., Huttner, S., Vavra, U. et al. (2009) Class I alpha-mannosidases are required for N-glycan processing and root development in *Arabidopsis thaliana*. *Plant Cell*, **21**, 3850–3867.
- Liu, Y., Wang, F., Zhang, H., He, H., Ma, L. and Deng, X.W. (2008) Functional characterization of the Arabidopsis ubiquitin-specific protease gene family reveals specific role and redundancy of individual members in development. *Plant J.* **55**, 844–856.
- Liu, Q., Zhang, C., Yang, Y. and Hu, X. (2010) Genome-wide and molecular evolution analyses of the phospholipase D gene family in poplar and grape. *BMC Plant Biol.* **10**, 117.
- Lorrain, S., Trevisan, M., Pradervand, S. and Fankhauser, C. (2009) Phytochrome interacting factors 4 and 5 redundantly limit seedling de-etiolation in continuous far-red light. *Plant J.* **60**, 449–461.
- Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.
- Maher, C., Stein, L. and Ware, D. (2006) Evolution of Arabidopsis microRNA families through duplication events. *Genome Res.* **16**, 510–519.
- Meyers, B.C., Kozik, A., Griego, A., Kuang, H. and Michelmore, R.W. (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. *Plant Cell*, **15**, 809–834.
- Moffat, C.S., Ingle, R.A., Wathugala, D.L., Saunders, N.J., Knight, H. and Knight, M.R. (2012) ERF5 and ERF6 play redundant roles as positive regulators of JA/Et-mediated defense against *Botrytis cinerea* in Arabidopsis. *PLoS ONE*, **7**, e35995.
- Nakano, T., Suzuki, K., Fujimura, T. and Shinshi, H. (2006) Genome-wide analysis of the ERF gene family in Arabidopsis and rice. *Plant Physiol.* **140**, 411–432.
- O'Malley, R.C. and Ecker, J.R. (2010) Linking genotype to phenotype using the Arabidopsis unimutant collection. *Plant J.* **61**, 928–940.
- Ossowski, S., Schwab, R. and Weigel, D. (2008) Gene silencing in plants using artificial microRNAs and other small RNAs. *Plant J.* **53**, 674–690.
- Paterson, A.H., Freeling, M., Tang, H. and Wang, X. (2010) Insights from the comparison of plant genome sequences. *Annu. Rev. Plant Biol.* **61**, 349–372.
- Pesaresi, P., Scharfenberg, M., Weigel, M. et al. (2009) Mutants, overexpressors, and interactors of Arabidopsis plastocyanin isoforms: revised roles of plastocyanin in photosynthetic electron flow and thylakoid redox state. *Mol. Plant*, **2**, 236–248.
- Remington, D.L., Vision, T.J., Guilfoyle, T.J. and Reed, J.W. (2004) Contrasting modes of diversification in the Aux/IAA and ARF gene families. *Plant Physiol.* **135**, 1738–1752.
- Reyes, F., Leon, G., Donoso, M., Brandizzi, F., Weber, A.P. and Orellana, A. (2010) The nucleotide sugar transporters AtUTr1 and AtUTr3 are required for the incorporation of UDP-glucose into the endoplasmic reticulum, are essential for pollen development and are needed for embryo sac progress in *Arabidopsis thaliana*. *Plant J.* **61**, 423–435.
- Riefler, M., Novak, O., Strnad, M. and Schumling, T. (2006) Arabidopsis cytokinin receptor mutants reveal functions in shoot growth, leaf senescence, seed size, germination, root development, and cytokinin metabolism. *Plant Cell*, **18**, 40–54.
- Rizza, A., Boccaccini, A., Lopez-Vidriero, I., Costantino, P. and Vittorioso, P. (2011) Inactivation of the ELIP1 and ELIP2 genes affects Arabidopsis seed germination. *New Phytol.* **190**, 896–905.
- Roschzttardtz, H., Seguela-Arnaud, M., Briat, J.F., Vert, G. and Curie, C. (2011) The FRD3 citrate effluxer promotes iron nutrition between symptomatically disconnected tissues throughout Arabidopsis development. *Plant Cell*, **23**, 2725–2737.
- Rossini, S., Casazza, A.P., Engelmann, E.C., Havaux, M., Jennings, R.C. and Soave, C. (2006) Suppression of both ELIP1 and ELIP2 in Arabidopsis does not affect tolerance to photoinhibition and photooxidative stress. *Plant Physiol.* **141**, 1264–1273.
- Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386.
- Rutter, M.T., Cross, K.V. and Van Woert, P.A. (2012) Birth, death and subfunctionalization in the Arabidopsis genome. *Trends Plant Sci.* **17**, 204–212.
- Schwacke, R., Schneider, A., van der Graaff, E., Fischer, K., Catoni, E., Desimone, M., Frommer, W.B., Flügge, U.I. and Kunze, R. (2003) ARAMEM-

- NON, a novel database for *Arabidopsis* integral membrane proteins. *Plant Physiol.* **131**, 16–26.
- Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M. and Van de Peer, Y.** (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **99**, 13627–13632.
- Strasser, R., Altmann, F., Mach, L., Glossl, J. and Steinkellner, H.** (2004) Generation of *Arabidopsis thaliana* plants with complex N-glycans lacking beta1,2-linked xylose and core alpha1,3-linked fucose. *FEBS Lett.* **561**, 132–136.
- Suzuki, M., Kamide, Y., Nagata, N. et al.** (2004) Loss of function of 3-hydroxy-3-methylglutaryl coenzyme A reductase 1 (HMG1) in *Arabidopsis* leads to dwarfing, early senescence and male sterility, and reduced sterol levels. *Plant J.* **37**, 750–761.
- Suzuki, M., Nakagawa, S., Kamide, Y., Kobayashi, K., Ohyama, K., Hashinokuchi, H., Kiuchi, R., Saito, K., Muranaka, T. and Nagata, N.** (2009) Complete blockage of the mevalonate pathway results in male gametophyte lethality. *J. Exp. Bot.* **60**, 2055–2064.
- Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. and Paterson, A.H.** (2008a) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.
- Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M. and Paterson, A.H.** (2008b) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954.
- Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L. and Vandepoele, K.** (2009) The flowering world: a tale of duplications. *Trends Plant Sci.* **14**, 680–688.
- Vandepoele, K., Simillion, C. and Van de Peer, Y.** (2002) Detecting the undetectable: uncovering duplicated segments in *Arabidopsis* by comparison with rice. *Trends Genet.* **18**, 606–608.
- Wang, H. and Deng, X.W.** (2002) *Arabidopsis* FHY3 defines a key phytochrome A signaling component directly interacting with its homologous partner FAR1. *EMBO J.* **21**, 1339–1349.
- Wang, J., Long, Y., Wu, B., Liu, J., Jiang, C., Shi, L., Zhao, J., King, G.J. and Meng, J.** (2009) The evolution of *Brassica napus* FLOWERING LOCUS T paralogues in the context of inverted chromosomal duplication blocks. *BMC Evol. Biol.* **9**, 271.
- Watkins, N.J. and Bohnsack, M.T.** (2012) The box C/D and H/ACA snoRNPs: key players in the modification, processing and the dynamic folding of ribosomal RNA. *Wiley Interdiscip. Rev. RNA*, **3**, 397–414.
- Weigel, M., Varotto, C., Pesaresi, P., Finazzi, G., Rappaport, F., Salamini, F. and Leister, D.** (2003) Plastocyanin is indispensable for photosynthetic electron flow in *Arabidopsis thaliana*. *J. Biol. Chem.* **278**, 31286–31289.
- Welchen, E., Hildebrandt, T.M., Lewejohann, D., Gonzalez, D.H. and Braun, H.P.** (2012) Lack of cytochrome c in *Arabidopsis* decreases stability of Complex IV and modifies redox metabolism without affecting Complexes I and III. *Biochim. Biophys. Acta*, **1817**, 990–1001.
- Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V. and Provart, N.J.** (2007) An 'Electronic Fluorescent Pictograph' browser for exploring and analyzing large-scale biological data sets. *PLoS ONE*, **2**, e718.
- Yao, M., Wakamatsu, Y., Itoh, T.J., Shoji, T. and Hashimoto, T.** (2008) *Arabidopsis* SPIRAL2 promotes uninterrupted microtubule growth by suppressing the pause state of microtubule dynamics. *J. Cell Sci.* **121**, 2372–2381.
- Yokoyama, A., Yamashino, T., Amano, Y., Tajima, Y., Imamura, A., Sakakibara, H. and Mizuno, T.** (2007) Type-B ARR transcription factors, ARR10 and ARR12, are implicated in cytokinin-mediated regulation of protoxylem differentiation in roots of *Arabidopsis thaliana*. *Plant Cell Physiol.* **48**, 84–96.
- Yun, J.Y., Tamada, Y., Kang, Y.E. and Amasino, R.M.** (2012) *Arabidopsis* tri-thorax-related3/SET domain GROUP2 is required for the winter-annual habit of *Arabidopsis thaliana*. *Plant Cell Physiol.* **53**, 834–846.
- Zhang, Y., Cheng, Y.T., Qu, N., Zhao, Q., Bi, D. and Li, X.** (2006) Negative regulation of defense responses in *Arabidopsis* by two NPR1 paralogs. *Plant J.* **48**, 647–656.

Publikation 3:

An easy-to-use primer design tool to address paralogous loci and T-DNA insertion sites in the genome of *Arabidopsis thaliana*



METHODOLOGY

Open Access

An easy-to-use primer design tool to address paralogous loci and T-DNA insertion sites in the genome of *Arabidopsis thaliana*

Gunnar Huep[†], Nils Kleinboelting[†] and Bernd Weisshaar^{*}

Abstract

Background: More than 90% of the *Arabidopsis thaliana* genes are members of multigene families. DNA sequence similarities present in such related genes can cause trouble, e.g. when molecularly analysing mutant alleles of these genes. Also, flanking-sequence-tag (FST) based predictions of T-DNA insertion positions are often located within paralogous regions of the genome. In such cases, the prediction of the correct insertion site must include careful sequence analyses on the one hand and a paralog specific primer design for experimental confirmation of the prediction on the other hand.

Results: GABI-Kat is a large *A. thaliana* insertion line resource, which uses in-house confirmation to provide highly reliable access to T-DNA insertion alleles. To offer trustworthy mutant alleles of paralogous loci, we considered multiple insertion site predictions for single FSTs and implemented this 1-to-N relation in our database. The resulting paralogous predictions were addressed experimentally and the correct insertion locus was identified in most cases, including cases in which there were multiple predictions with identical prediction scores. A newly developed primer design tool that takes paralogous regions into account was developed to streamline the confirmation process for paralogs. The tool is suitable for all parts of the genome and is freely available at the GABI-Kat website. Although the tool was initially designed for the analysis of T-DNA insertion mutants, it can be used for any experiment that requires locus-specific primers for the *A. thaliana* genome. It is easy to use and also able to design amplimers with two genome-specific primers as required for genotyping segregating families of insertion mutants when looking for homozygous offspring.

Conclusions: The paralog-aware confirmation process significantly improved the reliability of the insertion site assignment when paralogous regions of the genome were affected. An automatic online primer design tool that incorporates experience from the in-house confirmation of T-DNA insertion lines has been made available. It provides easy access to primers for the analysis of T-DNA insertion alleles, but it is also beneficial for other applications as well.

Keywords: *Arabidopsis thaliana*, T-DNA, Insertion mutants, Paralog, Primer design, GABI-Kat

Background

Arabidopsis thaliana is widely and very successfully used as a model organism in basic plant research. After the completion of its genome sequence in the year 2000 [1], several large mutant collections have been established. In most cases, T-DNA insertional mutagenesis mediated

by *Agrobacterium tumefaciens* has been used for the generation of knock out alleles for reverse genetic approaches [2,3]. The insertion of T-DNA in the plant genome occurs almost randomly [4-6], and different methods for the identification of insertion sites in specific lines have been established. The most frequently used method is based upon flanking sequence tags (FSTs). FSTs are short DNA sequences, which flank the T-DNA insertion site and contain genome sequence information adjacent to the insertion site. They are generated with PCR-based methods after digestion of genomic

* Correspondence: bernd.weisshaar@uni-bielefeld.de

[†]Equal contributors

Center for Biotechnology & Department of Biology, Bielefeld University, Universitaetsstrasse 25, D-33615 Bielefeld, Germany



DNA and adapter ligation [7]. Their sequences can be compared to the *A. thaliana* genome sequence using BLAST [8] to predict the insertion position(s) of the T-DNA in a given line. The GABI-Kat collection is the world's second largest FST-based T-DNA insertion line collection for *A. thaliana* [9]. The FST data along with insertion site predictions and information about confirmed T-DNA insertion alleles is accessible at SimpleSearch, which is the user interface to the database at the website of the project [10]. GABI-Kat lines can be accessed via SimpleSearch, and links to the stock centres are provided if the line that contains the relevant allele has been donated to the American and/or European stock centres for *A. thaliana* seeds [11]. In case of direct orders at GABI-Kat, predicted insertions are requested rather than simply seed of GABI-Kat lines. Upon a user request, T2 plants of the respective line are grown and the insertion site prediction is confirmed at GABI-Kat by PCR with an insertion site-specific primer and a T-DNA border primer, followed by sequencing of the amplicon. The experimental data for confirmed insertion alleles is presented on the SimpleSearch website [12]. This includes the amplicon sequences as well as the sequences of the primers used in the confirmation PCR. A more detailed overview on the features of the SimpleSearch site is summarised in [9].

A major problem during the FST-based insertion site prediction in T-DNA insertion lines occurs when the FST sequence cannot be assigned unambiguously to a *single* specific locus in the *A. thaliana* genome. Such events are inevitable, because even in a small genome like the one from *A. thaliana* only about 10% of all genes encode unique proteins. All other genes have at least one additional homologue [13,14]. One reason for the occurrence of homologues within the genomes of eukaryotes lies in genome duplication events leading to paralogous genes. This has already been studied in detail in the original *A. thaliana* genome sequence analysis [1], and searchable databases are available which support analyses of duplication events and paralogous gene families [15]. Beside genes (and without considering transposable elements), also non-genic sequences of the *A. thaliana* genome occur in higher copy numbers. Several mechanisms have been discussed for the duplication events (reviewed for example in [16]). Regardless of the exact mechanism, the ultimate result of duplication events is that even after evolutionary diversification of the duplicated sequences, larger stretches of similar sequences occur at different positions in the *A. thaliana* genome. In this article we will refer to regions with more than one copy of similar DNA sequences in the genome as "paralogous regions", regardless if genic or non-genic regions are concerned. In this sense, we will also use the term "paralog" for individual regions in paralogous regions, even if those regions are non-genic and if the

genetic origin (i.e. duplication event) of the respective region is not clear.

In all large FST-based T-DNA insertion line collections, the FSTs have so far been used to predict a single locus in the genome as the corresponding insertion site. If the locus is located within a paralogous region, the prediction and the decision for one of the paralogs is error-prone. However, given that the sequences of the paralogous loci are known, the confirmation process at GABI-Kat, which considers the DNA sequence of the confirmation amplicon, is able to resolve ambiguities concerning the correct insertion position in most cases. Only if the paralogous regions contain (almost) identical sequences a definite assignment to a single locus is not possible.

We present data from example cases in which the correct insertion locus was identified only after PCR-based confirmation using optimised primers, even though FST-based insertion site prediction was unable to assign a unique best-fitting locus. Insertion site predictions were redone using the TAIRv10 genome sequence and BLAST, and multiple predictions derived from single FSTs were combined into "paralog groups". When attempting to confirm a prediction from such a group, specific primers (as far as possible) unique for relevant insertion sites were designed. We developed a primer design method that identifies possible primers using a multiple alignment, which enables the discrimination between the different paralogous regions. An optimised, easy to use version of the tool is available on the website of GABI-Kat and allows users to design primers at their own locus or genome position of interest.

Results and discussion

FSTs and T-DNA insertion site predictions

The GABI-Kat database contains insertion site predictions from about 135,000 FSTs, which were generated for the 93,504 lines in the T-DNA insertion line collection. During the generation of GABI-Kat FSTs, genomic DNA of individual T1 plants was digested with *Bfa*I, adaptors were added, and fragments containing T-DNA borders as well as sequences of plant origin next to the T-DNA were amplified with a T-DNA- and an adaptor-specific primer [7]. The length of the resulting amplicons is dependent on the position of the *Bfa*I recognition sites in the genome relative to the insertion site. In case of more than one T-DNA insertion in a given line, more than one amplicon might be generated in a single reaction. Due to different sizes of these amplicons, an insertion corresponding to a longer amplicon is measured at the tail of the FST sequence, and an insertion corresponding to a shorter amplicon at the head of the FST sequence. Usually the shorter amplicon causes a stronger signal because of higher fragment abundance after PCR. We refer to these

FSTs, which allow to correctly predict several insertion sites in one line from different regions of one FST, as “composite FSTs”. Such FSTs have also been described for the SALK collection [17].

We often observed that GABI-Kat FSTs contain sequence parts from borders of two distinct T-DNA insertions present in a single line. When addressing insertion site predictions in paralogous regions of the genome, predictions from “composite FSTs” had to be considered as well because they share the feature “additional BLAST hit from one FST”. The optimised analysis pipeline (see below) that has been established at GABI-Kat detects, in addition to paralogous hits, also hits from “composite FSTs”. The additional predictions derived from GABI-Kat FSTs by using this optimised pipeline have been made available with the GABI-Kat database release No 27 [12].

Initially, the insertion predictions were deduced from the FSTs in a 1-to-1 relation. Only the best BLAST hit from a given single FST was evaluated for the prediction of a single insertion site [9,18]. In order to address paralogous regions of the genome, we recalculated the insertion site predictions for all FSTs. For this new assessment, a 1-to-N relation of FST to insertion site predictions was implemented in the internal GABI-Kat database. To be able to filter for the most relevant predictions, three categories (designated 0, 1 and 2) were defined and assigned to the different types of insertion predictions deduced from a single FST based on the BLAST e-value. Category 0 was assigned to the prediction deduced from the best BLAST hit. This was the one that had been selected as the only prediction (1-to-1) before the extended analysis (1-to-N) was performed. Additional predictions from the evaluated FST were assigned to category 1 if the BLAST e-values were lower than $1e-3$, and to category 2 if the e-values were $1e-3$ or higher (for details see Methods). The additional predictions were taken into account during the confirmation process at GABI-Kat if necessary. This was especially important if the e-values of the BLAST hits for a given FST region were highly similar or even identical because paralogous genomic loci were affected. Only one of the insertion site predictions derived from a single region of an FST corresponds to the correct insertion locus. Details about the results from the “1-to-N” type FST evaluation and insertion site prediction are listed in Table S1, which is included in the document “Additional file 1”.

We observed that the prediction of category 0 could be wrong due to small errors in the FST sequence, even if the BLAST analysis results in a unique best hit. Consequently, analysis of only this locus would have made confirmation impossible. The access to several BLAST hits from one FST region allowed creating groups of paralogous insertion site predictions of categories 0 and 1, which were derived from subsets of the FSTs of the

respective line. In total, about 11,000 paralog groups were detected in the GABI-Kat FST dataset. If a paralogous prediction was addressed for confirmation experimentally, several predictions in the respective group were analysed during the confirmation process, if necessary. Until now, more than 1,200 groups with paralogous predictions in the GABI-Kat collection have been solved experimentally. If a prediction other than the best prediction has been confirmed, this prediction was made available in SimpleSearch in addition to the “category 0 prediction” which was included anyway.

Primers in paralogous regions of the genome

Even when most parts of paralogous sequences are highly similar or even identical, the individual sequences often differ at certain positions. Based upon sequence alignments of the genomic DNA sequences of the individual paralogs in groups of paralogous predictions, we developed a primer design algorithm that allows designing specific primers (see Methods). Uniqueness for the individual paralog was preferably constructed into their 3'-ends. Such primers can enable the determination of the correct insertion site prediction by PCR and sequencing, even if only one base pair differs in the paralogous regions surrounding the set of paralogous insertion site predictions. A direct comparison of PCR results with the different paralog-specific primers allows the discrimination between the paralogs, sometimes taking into account that mispriming usually leads to weaker PCR products.

In addition to paralogous regions in the genome, random mispriming sites in the genome might occur for primers. In our experience and with our PCR conditions, even short sequence stretches at the 3'-end of primers can lead to unspecific PCR products (see [19] and references therein). We have regularly observed examples of primers, which were able to amplify unspecific PCR products when only 12 bases of their 3'-end had a perfect match in the genome. For example, in GABI-Kat line 011B05 we tried to confirm the predicted insertion at position 51,137 on chromosome 3. The primer that was used for this purpose (5'-CTCAATTTATGTGT GACTGCAAGC-3') had the unique, perfect annealing site from position 50,794 to 50,817 on chromosome 3. Unexpectedly, we observed an amplicon of roughly 1.3 kb. BLAST analysis of the sequence of this amplicon resulted in a hit in the gene *At4g33170* with a BLAST e-value of 0.0 and a derived T-DNA insertion site at position 15,997,766 on chromosome 4. Analysis of the primer sequence showed a perfect match of the last 12 bp at the positions 15,996,481 to 15,996,492 on chromosome 4. The insertion in the line 011B05 was subsequently confirmed with an *At4g33170*-specific primer, essentially by using the “wrong” confirmation sequence as an FST for

insertion site prediction. More extreme examples of mis-priming occur in rare cases. This is taken into account in our primer design by minimizing the number of possible 12 bp-matches within the genome (see Methods).

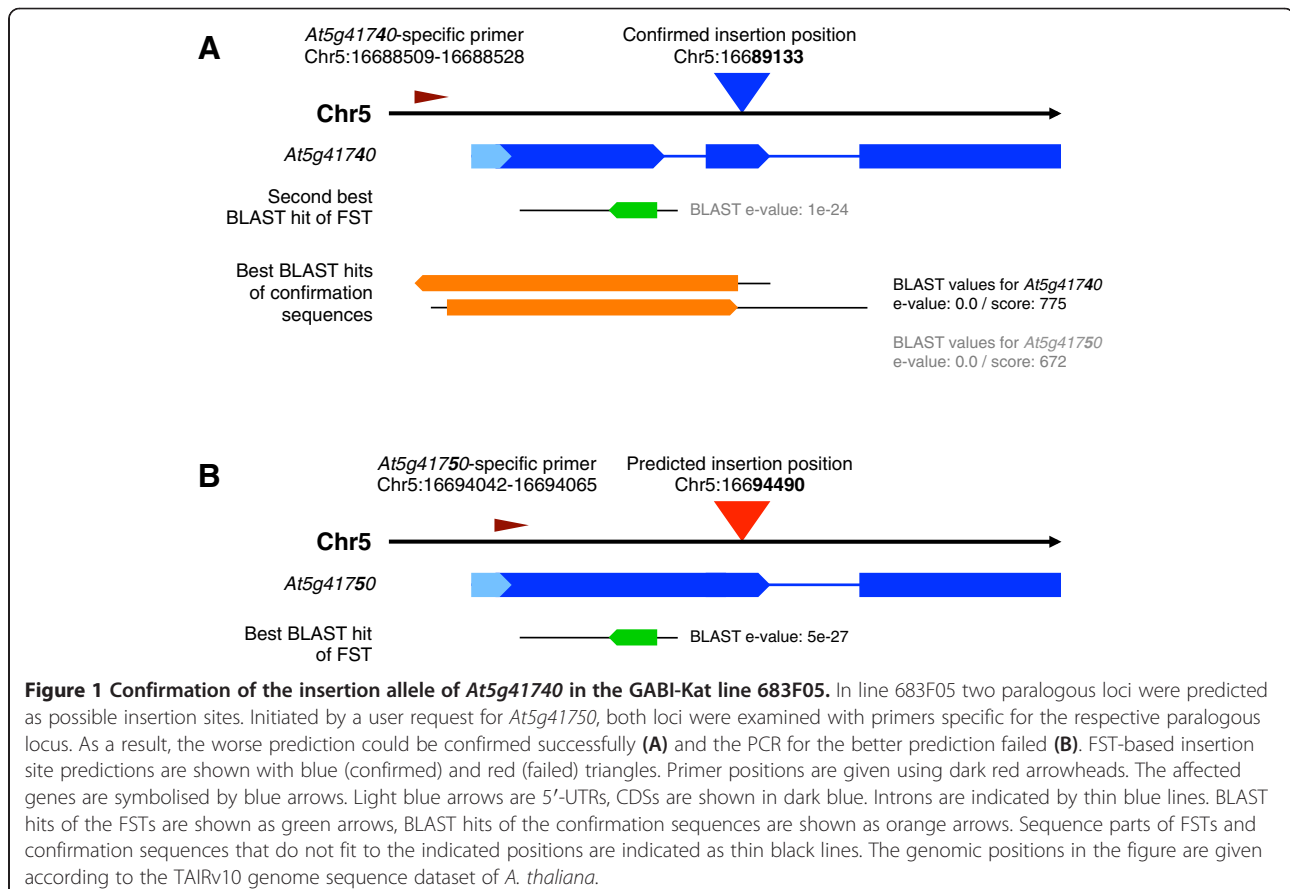
Application of the paralog primer method

An example that illustrates the advantages of the paralog primer method is the confirmation of the insertion in *At5g41740* in the GABI-Kat line 683F05 (see Figure 1). The best insertion site prediction and the only one that had been available in the “1-to-1 prediction dataset” in this line was *At5g41750* with a BLAST e-value of $5e-27$. This gene is annotated to encode a “disease resistance protein (TIR-NBS-LRR class) family” in TAIRv10. After newly calculating the insertion site predictions and setting up the “paralog groups”, the second best prediction for the same FST was *At5g41740* with a BLAST e-value of $1e-24$ and the same annotation. The experimental analysis with paralog specific primers designed using the tools described above resulted in an amplicon with the *At5g41740*-specific primer and no product with the *At5g41750*-specific primer. BLAST analysis of the sequence of the confirmation amplicon resulted in fully aligning sequences for both genes. However, when

comparing the score values, *At5g41740* reached 775 while *At5g41750* ended up with 672. Manual inspection of the alignments confirmed a few SNP positions that distinguish the two loci with *At5g41740* representing the correct locus. The confirmed insertion in *At5g41740* is available in SimpleSearch with a search for the respective AGI gene/locus code, and the data from the experimental analysis as well as a link to the stock centre NASC are displayed.

Another example is the confirmation of the insertion in *At1g07930* in the line 902F05. In this line the best predictions were *At1g07920* and *At1g07940* with a BLAST e-value of $1e-95$ in both cases. A third and slightly worse prediction for the same FST was *At1g07930* with a BLAST e-value of $7e-91$. Only for *At1g07930* a PCR product could be obtained with paralog-specific primers. BLAST analysis of the sequence of the amplicon confirmed *At1g07930* as the correct paralog via the score values similar to the example above (data available at the SimpleSearch website for line 902F05).

Besides examples of GABI-Kat lines with second- or third-best insertion site predictions being confirmed, there are several cases of lines which have two or more predictions with identical reliability according to the



BLAST e-values of the FSTs. Primers designed using the tools described above allowed the determination of the correct insertion locus in these lines. Examples are the GABI-Kat lines 583H04 and 742E06. In line 583H04, *At1g29350* was identified as a wrong prediction and the insertion in *At1g29370* was confirmed. In line 742E06, *At2g38210* was the wrong prediction and *At2g38230* was confirmed.

Access to the easy-to-use primer design tool

In order to offer public access to the primer design algorithms developed at GABI-Kat, we implemented an easy-to-use tool into SimpleSearch that includes the paralog-specific design if necessary. The design method is chosen as described in Methods and an overview on the primer design process is shown in Figure 2. Details on the selection of suitable primers using the paralog-specific design are summarised in Figure 3. The publicly available primer tool was implemented within the visualisation part and displays the location of the designed

primers. It can be accessed directly with the URL [20] or via the menu on the GABI-Kat website [10]. Insertions found in SimpleSearch also provide a link to the primer design for their respective position and a button in the visualisation allows quick access to the primer design for the currently selected position in the genome (Figure 4). It differs from previously available tools for the analysis of paralogous regions in a number of important aspects. Other tools, for example the very useful tool Primer-BLAST [19], checked the redundancy of the combination of both primer annealing sites for the amplicers defined by a primer pair. It also uses MegaBLAST of the complete target zone to the genome sequence of the addressed organism to avoid primer design in redundantly matching parts of the target zone. In contrast to this, our tool checks the 3'-ends of every single primer for redundancy, which is essential for the analysis of T-DNA insertions because in this experimental setup only one genome-specific primer is used. Primer-BLAST uses a BLAST of the complete target zone sequence and

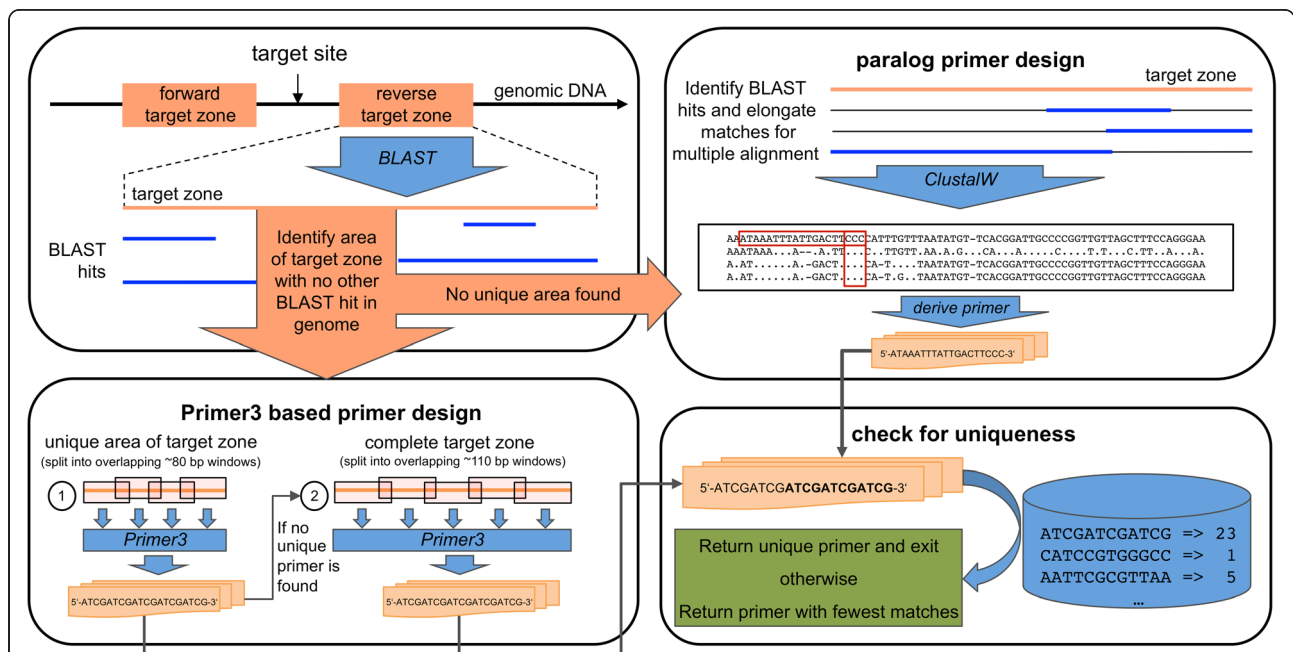
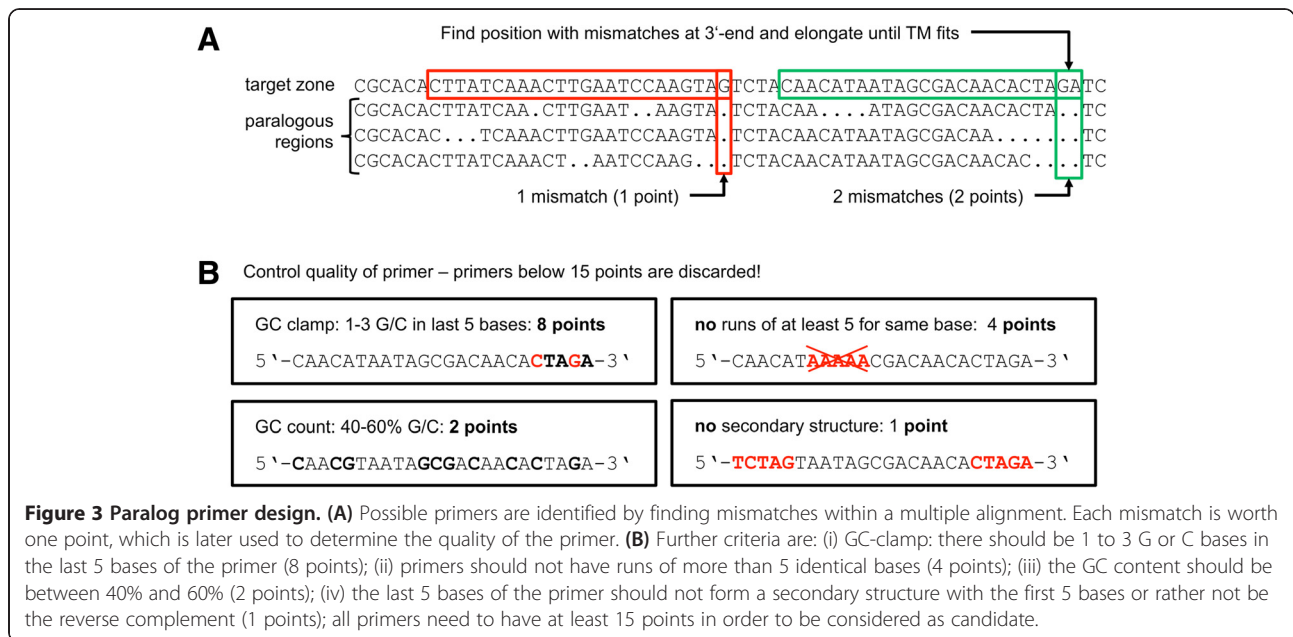


Figure 2 Overview on the primer design tool. Initially, each of the two target zones with a distance to the chosen target site (distance and target site defined by the user) is examined by a BLAST vs. the *A. thaliana* genome sequence. If there is an area that has no other BLAST hit somewhere in the genome, the Primer3-based approach is used (bottom left box), otherwise the paralog primer design is used (upper right box). A number of candidate primers is designed in both approaches which are then checked for uniqueness (bottom right box). If a unique primer with no additional 12 bp-hit at the 3'-end in the genome is found, the primer design is stopped and the primer is returned as a result. Otherwise the primer with the fewest matches is returned. When designing additional primers, the next best primers are returned. The Primer3-based primer design (bottom left box) uses multiple runs of Primer3 in overlapping windows and altering temperatures to generate a large set of candidate primers. First, only the unique area of the target zone with no additional BLAST hit in the genome is considered. If no unique primer is found, the process is performed again with the complete target zone. The paralog primer design first creates a multiple alignment with all sequences showing a BLAST hit to the target zone using ClustalW. The algorithm searches for mismatches in the multiple alignment (see Figure 3 for details). To reduce the runtime of ClustalW for sequences with many hits, the target zone is also split into overlapping windows and alignments are computed separately (not shown in Figure). The primer sequences shown in the figure are to be regarded as example sequences that cannot fit to *all* features of the scheduled workflow.



optimisation of the BLAST alignments by the Needleman-Wunsch global alignment algorithm to perform a specificity check that deselects primer pairs with amplicons on other targets than the submitted template [19]. Our approach to rank primer candidates uses a pre-computed index of all occurrences of 12 bp sequences in the *A. thaliana* genome and considers the last 12 bp of a primer candidate. This 12 bp stretch was in our experience the crucial part of the primer. For using the tool, a genomic nucleotide position central to the locus to be addressed (designated “target position”) must be selected. Upon starting the tool, primers are automatically designed around this target position with a default minimal and maximal distance of 300 to 800 bp to the target position on each side. We refer to this sequence range surrounding the target position as “target zone”. The distance to the target position can be changed to values between 100 to 1500 bp with a minimum range on each side of the target position of 100 bp. The larger the target zone, the better are the chances to obtain a unique primer. For the primers, the default annealing temperature has been set to 60.5°C, but it can also be set by the user to a value between 50 and 72°C. The primers can either be used in combination with T-DNA border primers in order to confirm a T-DNA insertion of interest, or simply to create amplicons from the targeted genomic locus. If the automatically designed primers are not acceptable to the user for some reason, the design tool can be executed repeatedly to acquire further primer combinations.

As an additional usability feature, the tool determines and reports the size of the amplicon with respect to the

pseudochromosome sequence, and presents a summary of information related to genotyping insertion alleles if a (predicted) insertion site is spanned by the amplicon. The PHP code of the tool is available upon request.

Conclusions

We describe the primer design procedure that has been used successfully and in large scale for confirmation of T-DNA insertion alleles in the GABI-Kat project. Since 2007 [21] users can access the sequences of experimentally proven confirmation primers for confirmed insertion alleles via SimpleSearch. Now, the primer design procedure established at GABI-Kat has been integrated into the publicly available SimpleSearch interface. The tool can be useful for confirming T-DNA insertion alleles, including those from SALK or other insertion mutant collections. At GABI-Kat, usually only one insertion is confirmed per line. After this first confirmation, the lines are donated to NASC and can further on only be ordered from there. Access to the GABI-Kat primer design tool might therefore help in the analysis of additional insertions, which are listed as predictions in SimpleSearch. Moreover, the tool allows easy design of amplicons for the genotyping of insertion alleles because the amplicons spanning the insertion site differentiate between the wt allele (amplicon produced) and the insertion allele (no amplicon; see [14]). The tool presented in this work differs from the already available tools in several aspects, as discussed above. Also, we simplified the primer design process by providing an easy-to-use user interface, which only requires several mouse clicks and no copy-paste of target sequences. Furthermore, existing

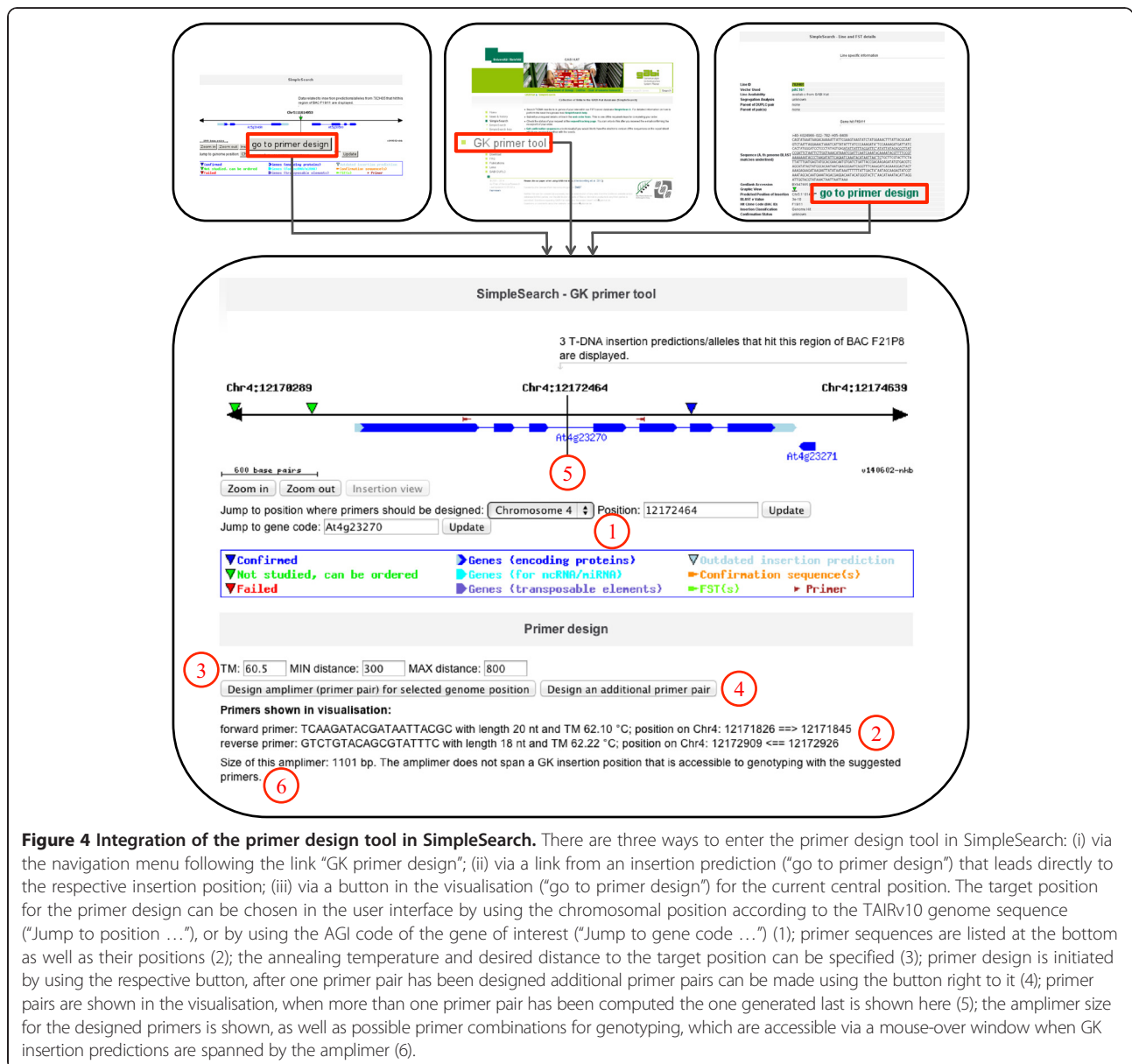


Figure 4 Integration of the primer design tool in SimpleSearch. There are three ways to enter the primer design tool in SimpleSearch: (i) via the navigation menu following the link “GK primer design”; (ii) via a link from an insertion prediction (“go to primer design”) that leads directly to the respective insertion position; (iii) via a button in the visualisation (“go to primer design”) for the current central position. The target position for the primer design can be chosen in the user interface by using the chromosomal position according to the TAIRv10 genome sequence (“Jump to position ...”), or by using the AGI code of the gene of interest (“Jump to gene code ...”) (1); primer sequences are listed at the bottom as well as their positions (2); the annealing temperature and desired distance to the target position can be specified (3); primer design is initiated by using the respective button, after one primer pair has been designed additional primer pairs can be made using the button right to it (4); primer pairs are shown in the visualisation, when more than one primer pair has been computed the one generated last is shown here (5); the amplicon size for the designed primers is shown, as well as possible primer combinations for genotyping, which are accessible via a mouse-over window when GK insertion predictions are spanned by the amplicon (6).

tools usually require the definition of several parameters by their users for the design of primers, which is often laborious and confusing for the users, especially when the underlying algorithms are unknown to them. Our tool ensures most convenient primer design, because it only requires the absolute minimum of parameter definition. This is mainly the genome position to be addressed which is easily accessible through GenBank or SimpleSearch or even already known. In addition, the distance to the target position to be considered for primer design and a value for the desired melting temperature of the primers is required. A main advantage is that problems with difficult genomic positions are taken care of automatically. With the new tool we hope to contribute to

the simplification of the analysis of T-DNA insertions as well as of other PCR-based applications in *A. thaliana*.

Methods

Plant growth conditions as well as molecular biological methods used for generation of the data used in this study have been described in [3]. General database aspects have been described in [9,21]. All FSTs generated in GABI-Kat are publicly available through ENA/GenBank and SimpleSearch.

Terminology

We use the term “amplicon” to refer to the DNA fragment that has been physically formed after PCR and

which can be sequenced. The term “amplimer” refers to the theoretical construct, which consists of a primer pair (located on opposite strands and with their 3′-ends directed towards each other) and the corresponding source sequence. For example, one can construct several amplimers addressing a predicted T-DNA insertion site, but only the primer pair that is based on correct predictions and/or assumptions about the configuration of the fusion of T-DNA to genomic DNA of the studied insertion allele will allow successful formation of an amplicon.

Categorisation of insertion sites

The evaluation of FSTs to predict insertion sites was based on hits (weighted sequence similarities) generated by BLAST [8] and by using the TAIRv10 genome sequence and annotation dataset [22]. Initially, only the best BLAST hit for each FST had been stored [18]. To address paralogous predictions, and also “composite FSTs” that contain sequence parts derived from at least two independent insertion sites, we changed the internal GABI-Kat database to allow storing several insertion site predictions for each single FST. The new predictions were categorised based on the e-value of the BLAST hit. Category 0 is assigned to the prediction deduced from the best BLAST-hit as long as its e-value is below (better than) $1e-3$. “Best” hits with larger (less significant) e-values were assigned to category 2. There can only be one best hit for each FST, and this hit is selected by using the top-of-the-list hit in the BLAST output, which results in the prediction of category 0. Note that the best hit might end up at the top position by chance if the respective FST region hits several parts of the genome with identical e-values and scores. If there are additional hits from the same FST with e-values below $1e-3$, the deduced predictions are classified as category 1. If different regions of the same FST have hits in different parts of the genome, these regions are handled individually to cover the cases of several insertions being deduced from one composite FST. From each of these regions (and in addition to the single category 0 prediction for the complete FST) a maximum of 3 hits are used to produce predictions of category 1; further BLAST hits are ignored. This restriction is necessary to reduce the amount of lab work caused by FST regions that are not only paralogous but repetitive. For further filtering, the deduced insertion sites (i) need to have a distance of 1000 bp to each other to be considered as a new insertion prediction, if they are closer to each other they are assigned to the same prediction, and (ii) are discarded if the e-value difference to the best hit for one FST region is larger than a factor of $1e10$. In general, our subsequent analyses considered only the predictions of categories 0 and 1.

Definition of paralog groups

For a systematic handling of insertion predictions that hit paralogous regions, we clustered them into groups using a hierarchical clustering approach for all predictions generated for a given GABI-Kat line (obviously this has been done for all lines). Starting with groups containing one prediction each, groups were combined if one of the following conditions hold true for each possible combination of predictions between the two distinct groups: (i) the prediction for different loci was based on the same part of the FST sequence (with a minimal overlap of 30 bp); (ii) the 400 bp of sequence next to the predicted insertion site have an identity of more than 79% to all members of the group. The clustering stopped when no further groups could be combined. All groups that contained at least two predictions for distinct insertions (i.e. they must display more than 1000 bp distance) were stored.

Primer design avoiding multiple annealing sites for the 3′-end

A primer was regarded as “unique” if its 12 bp-3′-end had only one hit in the genome sequence. The number of occurrences of each 12 bp sequence within the genome has been precomputed and stored in our database. By using this index, the number of possible matching positions for each primer can be identified easily by a simple and fast database query.

We have developed two methods for primer design. One is based on the widely used primer design tool Primer3 [23] with additional filtering, the other uses a self-developed algorithm that searches for mismatches within a multiple alignment of sequence-related target zones. For in-house confirmation at GABI-Kat, primers for insertions that are not located in paralogous regions are designed using the first method, while the second method is applied to insertions in paralogous regions. The public primer design tool works for all positions within the *A. thaliana* genome, is not dependent on (predicted) GABI-Kat insertion sites and automatically chooses the best method for the genomic locus addressed. In order to decide which method is suited best for the target zone in question, a BLAST of the sequences of this zone (that is, the sequences surrounding the target position limited by the value set for the distance to the target position) is performed against the *A. thaliana* genome sequence with an e-value cutoff of $1e-5$, which is high enough to detect hits of down to 24 bp. If there is a sequence part within the target zone that has no other hit in the *A. thaliana* genome and is at least 100 bp long, problems with paralogous regions should not occur and the Primer3-based method is used. If such a unique sequence part cannot be determined in the target zone sequence, the paralog primer design method is chosen.

Primer3-based method

In order to find primer candidates within a target zone, the part of the target zone without additional paralogous regions in the genome (identified during the decision which primer design method should be used) is used first. This part of the target zone sequence is divided into overlapping windows of at least 80 bp and for each of these windows a primer is designed by Primer3. Windows overlap with 30 bp to ensure that possible primers at the ends of the windows are considered as well. To further increase the number of possible primers the selected melting temperature is altered in steps of 0.4°C to maximally 1.2°C below or above the defined melting temperature. All primers are checked for uniqueness of their 3'-end as described above. As soon as a primer is detected that has only one 12 bp-hit within the genome, the primer design finishes successfully. If the initial search within the unique part of the target zone did not yield a result, the procedure is repeated within the whole target zone sequence with a window size of at least 110 bp and alternating melting temperatures in the same way as described above. If no unique primer could be found, the one with the fewest 12 bp-hits among all primers designed during the whole process is regarded as the best possible primer for this target zone.

Paralog primer method

In target zones that do have paralogous regions throughout their sequence somewhere in the genome, the Primer3-based approach often does not lead to satisfying results. Our algorithm first identifies all potentially paralogous regions within the genome by an initial BLAST with an e-value cutoff of 1e-5 and a minimum required length of 50 bp. All hits are elongated to fit the length of the target zone, and a multiple alignment is computed using ClustalW [24]. In order to reduce the runtime of ClustalW, a sliding window approach with overlapping windows of sizes around 220 bp (and an overlap of 30 bp) is used to compute the multiple alignments. In these multiple alignments, the algorithm searches for positions with a maximum number of mismatches to the sequence of the target zone. This position is defined as the 3'-end of a possible primer and is elongated to match the desired melting temperature. After that, the primer candidate is checked for GC-clamp (1–3 G/C in the last 5 nucleotides), base repeats (less than 5 identical nucleotides in a row), GC-content (between 40 and 60%) and secondary structures (last 5 nucleotides should not appear as reverse complement in the primer). The more of these conditions hold, the better the primer – primers not fulfilling some of these criteria are discarded (see Figure 3). The annealing temperature of the primer is computed using the same formula used in Primer3

(according to [25]) to achieve results comparable to those from the Primer3-based method:

$$T_m [^{\circ}\text{C}] = 81.5 - 11.6 + 0.41(\%GC) - \frac{600}{\text{length}}$$

A large number of candidates are generated and further checked for uniqueness as described above. If a unique primer is found it is returned as result. If this is not possible, the primer with the fewest matches among all examined primers is returned.

Additional file

Additional file 1: Table S1. The file contains a table (Table S1), which shows statistics about insertion site predictions in the GABI-Kat collection before and after the 1-to-N analysis of the FSTs. An explanation for the data presented in Table S1 is included in the file as well.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GH, NK and BW conceived and designed research, analysed and interpreted the data and wrote the manuscript. GH conducted wet-lab experiments. NK did database programming and bioinformatics. All authors read and approved the manuscript.

Authors' information

Gunnar Huep and Nils Kleinboelting are joint first authors.

Acknowledgements

The authors thank Andreas Kloetgen and Tina Zekic for their contributions to the implementation of the primer design, Yong Li, Mario Rosso, Prisca Viehoveer, the MPI for Plant Breeding Research and all former co-workers for their contribution to GABI-Kat, and Ute Buerstenbinder, Eliane Quittschau, Helene Schellenberg, Nina Schmidt, Andrea Voigt for technical assistance. The work described in this article is funded by the German Federal Ministry of Education and Research (BMBF) in the context of the German plant genomics program GABI (Förderkennzeichen 0313855). We acknowledge support of the publication fee by Deutsche Forschungsgemeinschaft and the Open Access Publication Funds of Bielefeld University.

Received: 23 June 2014 Accepted: 9 September 2014

Published: 13 September 2014

References

1. Initiative TAG: Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000, **408**(6814):796–815.
2. Alonso JM, Ecker JR: Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in *Arabidopsis*. *Nat Rev Genet* 2006, **7**(7):524–536.
3. Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B: An *Arabidopsis thaliana* T-DNA mutagenised population (GABI-Kat) for flanking sequence tag based reverse genetics. *Plant Mol Biol* 2003, **53**(1):247–259.
4. Szabados L, Kovacs I, Oberschall A, Abraham E, Kerekes I, Zsigmond L, Nagy R, Alvarado M, Krasovskaja I, Gal M, Berente A, Redei GP, Haim AB, Koncz C: Distribution of 1000 sequenced T-DNA tags in the *Arabidopsis* genome. *Plant J* 2002, **32**:233–242.
5. Li Y, Rosso MG, Ulker B, Weisshaar B: Analysis of T-DNA insertion site distribution patterns in *Arabidopsis thaliana* reveals special features of genes without insertions. *Genomics* 2006, **87**(5):645–652.
6. Kim S, Veena, Gelvin S: Genome-wide analysis of *Agrobacterium* T-DNA integration sites in the *Arabidopsis* genome generated under non-selective conditions. *Plant J* 2007, **51**(5):779–791.

7. Strizhov N, Li Y, Rosso MG, Viehoveer P, Dekker KA, Weisshaar B: **High-throughput generation of sequence indexes from T-DNA mutagenized *Arabidopsis thaliana* lines.** *BioTechniques* 2003, **35**(6):1164–1168.
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–410.
9. Kleinboelting N, Huep G, Kloetgen A, Viehoveer P, Weisshaar B: **GABI-Kat SimpleSearch: new features of the *Arabidopsis thaliana* T-DNA mutant database.** *Nucleic Acids Res* 2012, **40**:D1211–D1215.
10. **GABI-Kat Website.** [<http://www.gabi-kat.de>].
11. Scholl RL, May ST, Ware DH: **Seed and molecular resources for *Arabidopsis*.** *Plant Physiol* 2000, **124**(4):1477–1480.
12. **SimpleSearch.** [<http://www.gabi-kat.de/simplesearch.html>].
13. Armisen D, Lecharny A, Aubourg S: **Unique genes in plants: specificities and conserved features throughout evolution.** *BMC Evol Biol* 2008, **8**:280.
14. Bolle C, Huep G, Kleinbolting N, Haberer G, Mayer K, Leister D, Weisshaar B: **GABI-DUPLO: a collection of double mutants to overcome genetic redundancy in *Arabidopsis thaliana*.** *Plant J* 2013, **75**(1):157–171.
15. Ding G, Sun Y, Li H, Wang Z, Fan H, Wang C, Yang D, Li Y: **EPGD: a comprehensive web resource for integrating and displaying eukaryotic paralog/paralogue information.** *Nucleic Acids Res* 2008, **36**:255–262.
16. Rutter MT, Cross KV, Van Woert PA: **Birth, death and subfunctionalization in the *Arabidopsis* genome.** *Trends Plant Sci* 2012, **17**(4):204–212.
17. O'Malley RC, Ecker JR: **Linking genotype to phenotype using the *Arabidopsis* unimutant collection.** *Plant J* 2010, **61**(6):928–940.
18. Li Y, Rosso MG, Strizhov N, Viehoveer P, Weisshaar B: **GABI-Kat SimpleSearch: a flanking sequence tag (FST) database for the identification of T-DNA insertion mutants in *Arabidopsis thaliana*.** *Bioinformatics* 2003, **19**(11):1441–1442.
19. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL: **Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.** *BMC Bioinformatics* 2012, **13**(134):.
20. **GABI-Kat primer design.** [<http://www.gabi-kat.de/db/primerdesign.php>].
21. Li Y, Rosso MG, Viehoveer P, Weisshaar B: **GABI-Kat SimpleSearch: an *Arabidopsis thaliana* T-DNA mutant database with detailed information for confirmed insertions.** *Nucleic Acids Res* 2007, **35**:D874–D878.
22. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W, Mueller LA, Bhattacharyya D, Bhaya D, Sobral BW, Beavis W, Meinke DW, Town CD, Somerville C, Rhee SY: **The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant.** *Nucleic Acids Res* 2001, **29**(1):102–105.
23. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG: **Primer3—new capabilities and interfaces.** *Nucleic Acids Res* 2012, **40**(15):e115.
24. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23**(21):2947–2948.
25. Sambrook J, Fritsch EF, Maniatis T: *Molecular Cloning: A Laboratory Manual*. 2nd edition. New York, NY: Cold Spring Harbor Laboratory Press; 1989.

doi:10.1186/1746-4811-10-28

Cite this article as: Huep *et al.*: An easy-to-use primer design tool to address paralogous loci and T-DNA insertion sites in the genome of *Arabidopsis thaliana*. *Plant Methods* 2014 **10**:28.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



Publikation 4:

Evaluation of the structural features of thousands of T-DNA insertion sites indicates a double-strand break repair based insertion mechanism

1 **Evaluation of the structural features of thousands of T-DNA**
2 **insertion sites indicates a double-strand break repair based**
3 **insertion mechanism**

4

5 Nils Kleinboelting^{1†}, Gunnar Huep^{1†}, Ingo Appelhagen¹, Prisca Viehovever¹, Yong Li² and
6 Bernd Weisshaar^{1,*}

7

8 ¹ Center for Biotechnology, Bielefeld University, Universitaetsstrasse 25, D-33615
9 Bielefeld, Germany.

10 ² Department of Medicine IV, University Hospital Freiburg, Berliner Allee 29, D-79110
11 Freiburg, Germany.

12

13 * corresponding author
14 E-Mail: bernd.weisshaar@uni-bielefeld.de
15 Phone: +49 521 106 8720
16 Fax: +49 521 106 6423

17 † joint first authors.

18 E-mail addresses of all authors:

19 NK: nkleinbo@cebitec.uni-bielefeld.de

20 GH: ghuep@cebitec.uni-bielefeld.de

21 IA: ingo.appelhagen@jic.ac.uk

22 PV: viehoeve@cebitec.uni-bielefeld.de

23 YL: yong.li@uniklinik-freiburg.de

24 BW: bernd.weisshaar@uni-bielefeld.de

25 **Abstract**

26 Transformation by *Agrobacterium tumefaciens* is an important tool in modern plant research.
27 It involves the integration of a T-DNA initially present on a plasmid in agrobacteria into the
28 genome of plant cells. The process of attachment of the agrobacteria to plant cells and the
29 transport of the T-DNA into the cell and further to the nucleus has been well described.
30 However, the exact mechanism of integration into the hosts DNA is still unclear, although
31 several models have been proposed. During confirmation of T-DNA insertion alleles from the
32 GABI-Kat collection of *Arabidopsis thaliana* mutants, we have generated about 34,000
33 sequences from the junctions between the inserted T-DNA and adjacent genome regions.
34 Here, we describe the evaluation of this dataset with regard to existing models for T-DNA
35 integration. The results indicate that integration into the plant genome is mainly mediated by
36 the endogenous plant DNA repair machinery. The observed integration events showed
37 characteristics highly similar to repair sites of double-strand breaks with respect to
38 microhomology and deletion sizes. In addition, we describe unexpected integration events,
39 such as large deletions and inversions at the integration site that are relevant for correct
40 interpretation of results from T-DNA insertion mutants in reverse genetics experiments.

41

42 **Keywords**

43 T-DNA integration, *Agrobacterium*, double-strand break repair, *Arabidopsis thaliana*

44

45 **Running Title**

46 **T-DNA insertion mechanism shares features with DSB repair**

47 **Introduction**

48 Transgenic plants, especially from dicotyledon species, are usually generated using
49 *Agrobacterium tumefaciens* mediated T-DNA transfer (Gelvin, 1998). The T-DNA is
50 originally a part of the *A. tumefaciens* tumor-inducing (Ti)-plasmid (Hoekema et al., 1983;
51 Zambryski et al., 1980). It is flanked by two border sequences of 25 base pairs (bp) in length.
52 These border sequences are similar to each other and were classified as “left border” (LB) and
53 “right border” (RB) (Dürrenberger et al., 1989; Tinland et al., 1994). The original T-DNA,
54 which is transferred by naturally occurring agrobacteria, contains sequences with tumour
55 inducing properties in plants (Van Larebeke et al., 1974). The genes relevant for these
56 properties encode enzymes that force the plant to produce opines, which the bacterium can
57 subsequently metabolise and use as source of carbon and nitrogen (Schell et al., 1979). When
58 T-DNAs are exploited as molecular tools, the tumour inducing and metabolism affecting
59 genes are removed from the T-DNA. Moreover, it is possible to substitute all the sequences
60 between LB and RB without perturbing the ability of the T-DNA to integrate into the plant
61 genome (Peralta and Ream, 1985; Wang et al., 1984). These features of the agrobacterial plant
62 transformation system have been utilised to transfer many different sequences of choice into
63 plant genomes. The T-DNA system has been applied in very diverse experimental setups and
64 also in plant breeding (O'Malley and Ecker, 2010; Que et al., 2014).

65 While much is known about the transport of the T-DNA into the plant cell and to the nucleus,
66 the exact mechanism of integration is still unclear. The journey of the T-DNA into the plant
67 genome starts in *A. tumefaciens*: upon sensing of phenolic compounds from the plant by the
68 two-component signal transduction system composed of VirA and VirG, the expression of
69 virulence (Vir) proteins is activated, which leads to the production of a single stranded copy
70 of the T-DNA (T-strand) (Magori and Citovsky, 2011). During this process, VirD1 and VirD2
71 function together as an endonuclease complex. Restriction of the T-DNA takes place within
72 the border sequence right before the recognition sequence 5'-CAGGATATATT-3', shortening

73 the LB and RB sequences that "travel" together with the T-strand at LB by 3 bp and at RB by
74 22 bp, respectively (Jasper et al., 1994).

75 After binding of VirD2 to the 5'-end of the T-strand, the resulting complex is transported to
76 the plant cell by a type IV secretion system, composed of VirB and VirD4 proteins (Gelvin,
77 2010; Pitzschke and Hirt, 2010). VirD2 stays attached to the 5'-end of the T-strand during the
78 whole process and is also thought to protect the 5'-end from exonucleases. The integration of
79 the 5'-end of the T-DNA into the genomic plant DNA was assumed to be usually precise
80 (Dürrenberger et al., 1989). Within the plant cell cytoplasm, the T-strand is covered with
81 several VirE2 proteins (Citovsky et al., 1989). In addition, the bacterial effector VirE3 can
82 also bind directly to VirE2 and mediate the transport of the T-strand to the nucleus by a
83 nuclear localization signal (Lacroix et al., 2005). After uncoating, the T-strand is thought to
84 be converted into double stranded T-DNA (dsT-DNA) before integration (Tzfira et al., 2004).
85 Just recently, a model for the formation of dsT-DNA was proposed (Liang and Tzfira, 2013).
86 T-strands can also form T-circles, transiently expressing the genes on the T-DNA but
87 rendering the molecule useless for integration (Rolloos et al., 2014; Singer et al., 2012).

88 Because none of the exported effectors could be shown to play a role in T-DNA integration,
89 host factors are most likely exploited for the integration process. This is supported by the
90 characterization of several *Arabidopsis thaliana* ecotypes and mutants deficient in T-DNA
91 integration (Anand et al., 2007; Mysore et al., 2000a; Mysore et al., 2000b; Nam et al., 1997;
92 Nam et al., 1998).

93 Double-strand breaks (DSBs), which are caused for example by external effects like ionising
94 radiation or stress/reactive radicals but also are created by endogenous endonucleases during
95 biological processes like replication or recombination, can lead to serious deficiencies if left
96 unrepaired (Yoshiyama et al., 2013). In yeast and bacteria, DSBs are mainly repaired via
97 homologous recombination (HR), which usually returns the broken DNA to its original state
98 without altering the nucleotide sequence. In plants HR plays a minor role in DSB repair; the

99 main pathway is non-homologous end-joining (NHEJ). NHEJ is associated with various
100 rearrangements of the target site (see Gorbunova and Levy, 1999 and references therein). This
101 includes deletions, target site duplications, incorporation of short filler DNA unrelated in
102 sequence to the broken region (designated "fillers"), and inversions. The main proteins
103 involved in NHEJ are heterodimers of Ku70 and Ku80, which bind to DNA ends and recruit
104 several accessory factors that modify the ends for further processing with DNA ligase IV,
105 which forms a heterodimer with XRCC4. For DSB repair by NHEJ a small microhomology of
106 up to five bp is required between the two broken ends. Annealing at these microhomologies
107 leads to overhanging 3'-ends resulting in smaller deletions or target site duplications,
108 depending on the position of the microhomologous regions relative to the end of the DNA and
109 on the size of the single strand overhangs of the DSBs. The broken ends can also be repaired
110 without any region of microhomology by simple ligation, creating no alteration of the target
111 site, but this is a rare event (Gorbunova and Levy, 1999). Incorporation of fillers during NHEJ
112 has been explained by the synthesis-dependent strand annealing (SDSA) mechanism
113 (Gorbunova and Levy, 1999). According to this model, a single-stranded open end of the DSB
114 invades another double-stranded part of the genome and forms a migrating replication bubble
115 with weak association to the template. After abortion of synthesis and subsequent repair, a
116 filler at the repair site can be identified.

117 Besides HR and NHEJ, there is a third, less well characterised pathway for DSB repair, which
118 uses substantial microhomology to repair DSBs independent of Ku70, Ku80, DNA ligase IV,
119 XRCC4, and Rad52 - the key protein required for homologous repair. This so-called
120 microhomology-mediated end joining (MMEJ) can be distinguished from NHEJ by the use of
121 larger microhomologous sequences during alignment of broken ends. These microhomologies
122 usually range between 5-25 bp and result in larger deletions than observed in NHEJ. MMEJ
123 has also been associated with more complex rearrangements such as translocations or
124 inversions (see McVey and Lee, 2008 and references therein).

125 The most prominent model for T-DNA integration involves NHEJ DNA repair, in which
126 DSBs within the plant genome are targets for the integration (Chilton and Que, 2003;
127 Salomon and Puchta, 1998; Tzfira et al., 2003). Only studies in yeast have shown that
128 integration can also occur via HR (van Attikum et al., 2003). Both pathways (HR and NHEJ)
129 are conserved in plants. Nevertheless, NHEJ seems to be the predominant strategy for T-DNA
130 integration in plants (Britt and May, 2003; Ray and Langer, 2002). This is also supported by
131 the fact that downregulation of XRCC4 (one of the key proteins required for NHEJ) increased
132 the number of stable transformations in *A. thaliana* and *Nicotiana benthamiana* whereas
133 overexpression led to decreased T-DNA integration. It was also proposed that VirE2
134 suppresses XRCC4 activity. In either way, DSB repair is delayed and therefore availability of
135 more DSBs should facilitate T-DNA integration (Vaghchhipawala et al., 2012). In a study that
136 provided evidence that T-DNA is integrated in DSBs as dsDNA, the DSBs in the host DNA
137 and in the T-DNA were created by a rare cutting restriction enzyme and precise ligation into
138 the restriction site was observed (Tzfira et al., 2003). During NHEJ, often a few nucleotides
139 are deleted at the insertion site (Weaver, 1995), which has also been observed for some T-
140 DNA insertions (Windels et al., 2003), but comparison to the size of typical deletions after
141 DSB repair showed the deletions after T-DNA integration to be much smaller (Kirik et al.,
142 2000; Orel and Puchta, 2003). Also, small areas of microhomology between the T-DNA and
143 the neighbouring plant genomic DNA have been found during an analysis of flanking
144 sequence tags (FSTs, Brunaud et al., 2002).

145 Besides the DSB repair based model for integration, two models for the integration of single-
146 stranded T-DNA have been proposed (Tzfira et al., 2004). In the single-strand-gap repair
147 model (SSGR) the single-stranded T-DNA binds with both borders to the genomic DNA by
148 microhomology within an existing gap that has been created by elongation of an existing nick
149 by an exonuclease. The other genomic DNA-strand is then removed by endo- and
150 exonucleases and repaired using the T-DNA as template. In the microhomology-dependant

151 model, areas of microhomology are used to initiate base-pairing of the T-DNA to the
152 unwound genomic DNA before a nick at the integration site is created in one strand. Pairing at
153 areas of microhomology then also takes place at the other border, the intact genomic strand is
154 also cut by an endonuclease and partly removed, followed by repair using the T-DNA as
155 template. A variable size deletion would always be the result for both of these models as well
156 as the presence of substantial microhomology for the second model. Single-stranded based
157 integration models have been initially preferred because it was also shown that the
158 transformation frequency for ssDNA is higher than for dsDNA in plant protoplasts
159 (Rodenburg et al., 1989). Therefore it was assumed that single stranded T-DNA is also
160 preferred for integration and that nicks in the ds plant DNA are used for the integration
161 (Gheysen et al., 1991; Mayerhofer et al., 1991).

162 It is widely accepted that T-DNA insertions within plant genomes occur such that the
163 sequences between the borders (LB and RB) are inserted at a random position in the genome
164 in single copy (Forsbach et al., 2003). But it is also known that more complex arrangements
165 of T-DNAs including vector backbone sequences from outside of the T-DNA itself may occur
166 at the insertion sites (see Krizkova and Hroudá, 1998 and references therein). Also, parts of
167 the *A. tumefaciens* chromosome sequences are carried over to the plant genome together with
168 T-DNA (Ulker et al., 2008). In contrast to a model where a single stranded T-DNA is
169 incorporated into the plant DNA, the ds model can also explain the integration of such
170 complex T-DNA inserts containing multiple T-DNAs linked together. This is due to the fact
171 that T-strands cannot recombine at their right borders when they are not in ds form (De Buck
172 et al., 1999; De Neve et al., 1997).

173 The integration sites within the plants' genomes are largely randomly distributed, although it
174 is still discussed if there are preferences for certain genomic regions. It has been shown that
175 there are preferences for transcription initiation sites and polyadenylation sites and regions
176 outside of centromeric regions, at least after selecting insertion events with a marker gene (Li

177 et al., 2006; Szabados et al., 2002). In contrast to this, it was described that the integration
178 sites are distributed within the genome in a completely random manner (Kim et al., 2007).
179 The discrepancy of these observations can be explained at least in part by a "selection bias"
180 caused by selecting T-DNA containing plants with an antibiotic resistance conferred by the T-
181 DNA. This bias would explain why insertions in transcriptionally inactive genomic regions
182 might be under-represented, because the antibiotic resistance gene expression might not be
183 sufficiently active. From a mechanistic point of view and by excluding "selection bias" the
184 distribution of integration sites seems to be random.

185 A special application for T-DNA insertions in plant genomes is the generation of large FST-
186 indexed insertion line collections. The collections are a resource for potential knockout
187 alleles, because a T-DNA insertion within or close to a gene usually perturbs this gene's
188 function. FSTs allow the prediction of the insertion site position of the T-DNA in specific
189 lines. The FSTs are generated by digestion of genomic DNA from many individual lines
190 followed by adapter ligation as well as PCR-based methods using T-DNA specific primers,
191 adapter primers, and subsequent sequencing of the PCR products (Alonso et al., 2003;
192 Strizhov et al., 2003; Thole et al., 2009). For *A. thaliana*, several T-DNA insertion line
193 collections exist, of which the SALK and GABI-Kat collections are the largest, comprising
194 about 88,000 and 72,000 indexed lines, respectively. Both collections focussed on analysing
195 junctions to genomic sequences at the LB, because it turned out empirically that the success
196 rate of FST production was higher. At least in part, this can be explained by the fact that T-
197 DNA integrations often appear in a complex LB-rb::rb-LB configuration that contains some
198 sort of internal fusion of the right parts of the T-DNA (that does not need to contain RB
199 sequences) and results in LB at both junctions to genomic sequences (Strizhov et al., 2003).
200 User access to the collections is provided through publicly available databases, containing
201 FST data. This allows users to search for insertion lines for their genes of interest (Alonso et
202 al., 2003; Kleinboelting et al., 2012). The lines can be ordered, either at common stock centres

203 (Scholl et al., 2000) or directly at the web pages of the collections. In case of the GABI-Kat
204 collection, the FST predictions for the insertion sites of the T-DNA are confirmed via PCR
205 with border primers and insertion-site-prediction-specific primers and sequencing, before
206 ordered lines are sent out to users. These confirmation sequences are in most cases of better
207 quality than the FSTs and allow an improved estimation of the exact insertion position. The
208 FSTs in SALK and GABI-Kat are mostly prepared at the LB side of the T-DNA (see above),
209 but for GABI-Kat lines there are additional FSTs from the RB side for a subset of the
210 collection.

211 In this study we present a detailed analysis of confirmation sequences obtained over the years
212 in the GABI-Kat project. For a quite large number of randomly selected lines, confirmation
213 sequences were created for both junctions of the insertion, which enabled a comprehensive
214 analysis of the size of deletions caused by the integration mechanism. Additionally, we
215 investigated similarities between the deleted regions and the T-DNA ends. Using large
216 numbers of reliable confirmation sequences, we analysed if the microhomology at the
217 observed T-DNA/plantDNA junction is a random sequence similarity effect or not.
218 Furthermore, we observed some unusual integration events where large parts of the genomic
219 DNA were deleted as well as rare cases where a part of the plant DNA was inverted at the
220 insertion site.

221

222

223 **Results**

224

225 **Size of RB and LB sequences remaining at T-DNA to genome junctions**

226 As a first step, we determined the number of bases that were lost from LB and RB of the
227 original T-DNA sequence, which initially includes two complete border sequences. The size
228 distribution of T-DNA border sequences detectable after integration is depicted in Figure 1.

229 For LB, 11,103 individual confirmation sequences that each correspond to exactly one
230 junction of an insertion were evaluated. The peak for LB was observed at position -3, with
231 17.7 % of all LB sequences showing a loss of 3 bp. In 72.3 % of all cases, 4 or more bases
232 were lost while for 10.0 % of the cases longer border sequences were found which lost less
233 than 3 bp. The largest fraction of these longer border sequences lost only 2 bp (3.2 %). In a
234 few cases, sequence from the vector neighbouring the T-DNA ("outside" of the border) was
235 integrated as well (5.6 %) with a median length of 13 bp of these additional sequences.

236 The number of cases available for analysis for RB was lower, because GABI-Kat FSTs were
237 mainly generated from the LB of the T-DNA and therefore fewer confirmation sequences at
238 the RB were available for analysis.

239 A total of 699 insertions were analysed for lost sequences of RB, and the peak for RB was
240 observed at position -22 (18.9 %). In 68.2 % of all RB cases 23 or more bases were lost. The
241 remaining 12.9 % represent cases with a longer RB sequence transferred into the *A. thaliana*
242 genome. In 8.3 % sequences from the vector were integrated with a median size of 21.5 bp.

243 The median position of the border cut for LB was -10 and -24 for RB, which corresponds to a
244 difference of -7 to the expected cut position for LB and -2 for RB, respectively, indicating that
245 LB gets shortened further than RB. Details on all evaluated confirmation sequences and their
246 cut positions can be found in Supplemental Information File SI1.

247 For both borders, the peak was located immediately upstream of the sequence pattern 5'-
248 CAGGATATATT-3'. We conclude that the main cut position precisely fits the position
249 defined previously, but also that there is considerable heterogeneity with respect to the actual
250 position of the cut which also happens to be located outside of LB and RB. The up to three
251 nucleotides that fit the original border sequence "outside" of the cut position might be
252 contributed by a random fit of sequences contributed by the pre-insertion site. The value for a
253 border one base longer than expected is approximately 25 % of the size of the peak for the
254 expected size. 25 % is roughly also the probability for a random matching base within the

255 plant genomic DNA. This also applies for further values of junctions with two, three or four
256 additionally fitting bases. The transfer of longer sequences containing vector sequence was
257 much more common than expected as a result from random sequence similarity. The loss of
258 sequences from the T-DNA "inside" of the cut position might be a consequence of the
259 integration mechanism.

260

261 **Microhomology and fillers at integration sites**

262 To examine the insertion sites in more detail, we determined the extent of microhomology of
263 sequences at individual T-DNA ends to the genomic sequence at the insertion site, and the
264 presence or absence of fillers, by sequence comparisons. Figure 2 shows an overview over the
265 size distribution of microhomology and fillers at integration sites. Supplemental Information
266 File SI1 contains the underlying data. In line with what has been mentioned above, we define
267 a filler as a DNA sequence present between the end of T-DNA sequence similarity of a given
268 insertion and the plant genomic DNA sequence at the insertion site that is not a direct
269 extension of either of the two. Overall, 11,802 sequences were analysed for both borders,
270 11,103 for the left border and 699 for the right border. Taken all sequences together (Figure
271 2A), fillers were detected on a regular basis in a total of 5,749 cases. Of the 11,802 insertions
272 studied, 48.7 % showed fillers of at least 1 bp, 2,209 (18.7 %) larger than 10 bp and 1,055 (8.9
273 %) were even larger than 20 bp.

274 The remaining 6,053 sequences without fillers were used to examine microhomology between
275 the individual T-DNA sequence ends and the adjacent plant genomic DNA. 5,351 (88.4 %) of
276 these insertions showed microhomology of at least 1 bp and the largest fraction of 4,525 (74.8
277 %) sequences had microhomology of up to 5 bp. The remaining 826 (13.6 %) sequences
278 showed microhomologies larger than 6 bp and the largest microhomology was 20 bp with 2
279 observed cases.

280 When separating the data for LB and RB (Figure 2B, Figure 2C), the size distributions were
281 found to be comparable. Fillers were observed in 48.3 % (5,368 cases) for LB and 54.5 %
282 (381 cases) for RB. The fillers were larger than 10 bp in 18 % / 29.8 % (2,001/208) and larger
283 than 20 bp in 8.4 % / 16.7 % (938/117) of cases. The mean filler for LB was 13/22.2 bp, the
284 median was 7/12 bp, indicating slightly larger fillers at RB.

285 When comparing microhomology sizes of LB and RB, 5,735 (51.7 %) sequences were
286 evaluated for LB and 318 (45.5 %) for RB. Among them, 5,081/270 (88.6 % / 85 %)
287 sequences showed microhomology of at least 1 bp, 4,296/229 (74.9 % / 72.0 %) of those was
288 between 1 and 5 bp and 785/41 (13.7 % / 12.9 %) were larger than 5 bp. The mean
289 microhomology is 3.18 bp for the LB and 3.48 bp for the RB, the median is 3 bp for both
290 borders.

291 Thus, both borders show a very similar pattern concerning microhomologies. In conclusion,
292 both borders show substantial microhomology to the *A. thaliana* genome sequence in roughly
293 half of the cases with similar size distribution. The remaining cases showed fillers, with
294 slightly larger fillers at the RB.

295

296 **Analysis of fillers**

297 To gain insights what could have caused the fillers, we determined the origin of the detected
298 fillers by comparing the sequences bridging the gap between T-DNA and genomic DNA (with
299 a small overlap) to possible origins. Table 1 gives an overview on the origin of the detected
300 fillers. 94.9 % of all best BLAST hits for the fillers identify sequences within the *A. thaliana*
301 genome. The second most abundant BLAST hits were found on T-DNA sequence (3.2 %).
302 Other hits were rare and can be explained as random sequence similarity due to the short
303 length of the sequences that had to be studied. Exceptions are sequences that might have been
304 introduced together with the T-DNA, namely sequences with hits to the *A. tumefaciens*
305 genome sequence or vector hits. Transfer of genomic DNA of *A. tumefaciens* along with T-

306 DNA has been described previously (Ulker et al., 2008). Here, we detected two fillers which
307 are larger than 20 bp that are derived from *Agrobacterium* chromosomes (see Supplemental
308 Information File SI2).

309 We examined the distance of the source sequence for the fillers to the location of the insertion
310 site containing the respective filler. Figure 3A shows the distribution of the distances to the
311 corresponding insertion position derived from the confirmation sequence. A large fraction of
312 all hits (36.3 % in comparison to 14.7 % as mean for the other four chromosomes) were found
313 on the same chromosome. Those fillers displayed a bias towards 1 kbp direct genomic
314 proximity to the insertion site (Figure 3B and 3C). Despite that bias for fillers within 1 kbp to
315 the insertion site, fillers from other genomic regions were randomly distributed within the
316 respective chromosome: without the high peak at close genomic proximity, the distance
317 distribution of positions resembles that of the difference of two equally distributed random
318 variables (one being the insertion site, the other the filler's source site). When all fillers were
319 evaluated, the distribution of hits on other chromosomes was equally distributed and scales
320 quite well with the relative size of the source chromosome. When fillers from insertions on an
321 individual chromosome were evaluated for their source location, there was also no convincing
322 bias for a certain source chromosome. (Figure 3D).

323 After identifying the origin of the filler sequences, we were able to check if there is
324 substantial microhomology between a given filler and the adjacent sequences at the fusions to
325 either T-DNA or the genome sequence (Figure 3E and 3F). Filler source sequences showed
326 small areas of microhomology to the sequence neighbouring the filler at the insertion site in
327 many cases: 22.8 % for the fusion to T-DNA and 25.8 % for the fusion to genome sequence.
328 This is approximately half of the rate of the T-DNA/genomic-DNA junction. The remaining
329 cases contained further “fillers of fillers” forming complex arrangements of sequences from
330 various sources, which we did not analyse further.

331 In conclusion, the large majority of fillers originate from the same chromosome or at least
332 from the *A. thaliana* genome, and from sequences available within the nucleus at the time of
333 integration (other T-DNA molecules and also sequences that "travel" together with the T-
334 DNA). There is a bias regarding the source of fillers towards the direct genomic
335 neighbourhood and the same chromosome of the integration site. The remaining filler sources
336 are randomly distributed throughout the genome. Fillers share some microhomology to their
337 neighbouring sequences at about half the frequency that was observed for T-DNA/genome
338 junctions without fillers. All data from the filler analysis is given in Supplemental Information
339 File SI3.

340

341 **Deletion and target site duplication of sequences from the pre-insertion site**

342 To determine the effect of the integration on the genomic sequences present at the pre-
343 insertion site, we generated confirmation sequences for both junctions (T-DNA/genome
344 intersections at both ends of the inserted T-DNA) for 1,319 individual insertions. These cases
345 were analysed for deletions and target site duplications, and the results are shown in Figure 4
346 with regard to the size of the deletions/duplications detected. Most of the insertions (93.3 %,
347 1,252 cases) displayed alterations of the pre-insertion site sequence between a target site
348 duplication of 19 bp (value of -19 in Figure 4) and a deletion of 99 bp (value of +99 in Figure
349 4). The remaining 6.7 % of cases displayed larger deletions or target site duplications. A total
350 of 80 insertions (6%) showed deletions of at least 100 bp and larger, and target site
351 duplications of at least 20 bp and larger were detected in 9 cases (0.7%).

352 Deletions of up to 100 bp were observed in 86 % of all examined insertions. The median of all
353 deletion sizes was 19 bp. We observed target site duplications in 89 cases (6.7 %) with a
354 median size of 4 bp. Only in 12 of 1,319 cases, neither a deletion nor a target site duplication
355 was found, indicating that deletions and target site duplications are a very common feature for

356 T-DNA integration with a strong bias towards deletions. A summary of all examined
357 insertions can be found in Supplemental Information File SI4.
358 In conclusion, integration without some alteration of the integration pre-insertion site in the
359 form of a either a deletion or a target site duplication is very rare. T-DNA insertion usually
360 leads to deletions of 1-50 bp with a smooth transition to target site duplications, while larger
361 alterations are the exception.

362

363 **Comparison with data from SALK lines**

364 To demonstrate that the observed insertion site structures are not specific for the T-DNA used
365 at GABI-Kat, we also analysed a number of insertions from the SALK collection (see
366 Supplemental Information Files SI5, SI6, and SI7). With regard to sizes of deletions and
367 target site duplications, the 69 analysed cases also showed a bias towards deletions of up to
368 100 bp (89.9%, 62 cases) and a median deletion size value of 19.5 similar to the result
369 obtained for GK insertions. The dataset for the analysis of the borders of inserted T-DNA,
370 microhomologies and fillers consisted of 880 SALK cases, 852 for the LB and 28 for the RB.
371 The SALK T-DNA sequences for RB and LB from pROK2 contain the identical recognition
372 sequence (see above) as found in the GK T-DNA, and we used this consensus to align the
373 position values. The peak of cutting sites in SALK insertions was at the expected position for
374 LB (-3) and RB (-22). However, for SALK LB cuts a smaller peak was located around -22 to
375 -24 bp, which resulted in an increased median value for the cut positions of -20 bp relative to
376 the expected site. The median RB cut position was -3.5 bp distance to the expected site, fitting
377 to the observations for GK RB cut positions.

378 Microhomology was observed in 413 (47 %) of the examined SALK cases with a median size
379 of 3 bp (3 for LB and 2 for RB). The median length of fillers was 8 bp (7 for LB and 17 for
380 RB). The origin of fillers was determined for 109 cases, all of them matched to the *A. thaliana*
381 genome sequence, with a small bias to the same chromosome (35 fillers). Frequency and size

382 of microhomologies or fillers and their origin matched to the values obtained from the
383 analysis of GK insertions, at least for the left border, where sufficient cases were analysed.
384 Taken together, the general picture of insertion site modifications is similar for GK and SALK
385 T-DNA alleles.

386

387 **Larger deletions**

388 We detected several larger deletions (31 insertions with deletions larger than 1000 bp, 8 of
389 them larger than 5 kb) at the integration site. Two of these deletions were analysed in more
390 detail, to confirm that the deleted region has really been removed and not translocated. In line
391 GK_144F03 a deletion of 15 kb on Chr3 between position 17,746,865 and 17,753,018
392 (according to TAIR10, see Methods) was predicted after evaluation of the confirmation
393 sequences. We sequenced DNA from homozygous T3 plants of this line using an Illumina
394 MiSeq, and mapped the reads (about 32.7x coverage) to the *A. thaliana*(Col) genome
395 sequence. The mapping result clearly confirmed the predicted large deletion because no reads
396 were mapping to the deleted region with respect to the wt sequence (see Figure 5 for a
397 coverage plot). Gene *At3g48070*, encoding a RING/U-box superfamily protein, is annotated
398 on Chr3 between position 17,750,711 and 17,752,761. The NGS re-sequencing result
399 confirms the complete removal of this gene from the genome sequence of GK_144F03,
400 although this is not apparent from the initial FST data.

401 In line GK_478B05 we predicted a deletion of approximately 6 kb on Chr3 between position
402 19,022,337 and 19,028,218. In this region the genes *At3g51230*, *At3g51238* and *At3g51240*
403 are annotated. *At3g51240* encodes Flavonone 3-Hydroxylase (F3H), an enzyme involved in
404 flavonoid biosynthesis. *F3H* mutants display the *transparent testa* (*tt*) phenotype, which
405 results from a reduced content of brown proanthocyanidins in the seed coat. The locus coding
406 for *At3g51240/F3H* has therefore been named *transparent testa 6* (*tt6*). We phenotypically
407 characterised homozygous offspring of GK_478B05 for the *tt6* phenotype and found all

408 phenotypic features expected for *tt6* (Figure 6). Therefore, the insertion allele of GK_478B05
409 does not contain a functional *At3g51240/F3H* gene. We conclude that also in this case the
410 region predicted to be deleted at the insertion site is absent from the genome of GK_478B05
411 (if homozygous for the insertion allele). A list of the genes (AGI codes) which were affected
412 by deletions identified in this study but that were not detectable from direct FST evaluation is
413 included in the supplements (Supplemental Information File SI8).
414 Detection of deletions larger than a few 100 bp were only possible if FSTs for both junctions
415 of the insertion had been available. This availability was a result of the FST generation
416 procedure of GABI-Kat which involved two PCR reactions with primers differing at the 3'
417 terminal base, which by chance also provided FST for both termini from one insertion in
418 some cases. Thus, the real frequency of larger deletions might be higher than our numbers
419 indicate.

420

421 **Inversions at integration site**

422 In some cases, we observed a seemingly incorrect orientation of the confirmation sequence
423 relative to the T-DNA. Upon detailed analysis of these cases, it turned out that genomic
424 sequences at the insertion site were reversed. These inversions create another breakpoint close
425 to the insertion position. Detection of such inversions were very rare, but similar to larger
426 deletions we cannot exclude that they are happening more often and are larger than the ones
427 we observed. Due to our preferred primer distance to the insertion site, and because we
428 (obviously) focus on correctly oriented primers, the confirmation amplicon can only form if a
429 given inversion is smaller than the distance of the primer to the insertion position.
430 Nevertheless, we found 10 inversions of different sizes in our confirmation sequences. Table
431 2 summarises the data and the location and sizes of the inversions.

432

433

434 **Discussion**

435 GABI-Kat generated a large publicly available dataset of sequence-characterised T-DNA
436 insertions with high sequence quality and accurate insertion site positions in *A. thaliana*. We
437 have analysed this data to obtain further evidence for or against the available hypothesis on
438 the mechanism of the T-DNA insertion process. The most popular theory assumes that T-
439 DNA integrates as a double-stranded DNA molecule into double-strand breaks and utilises
440 plant repair pathways for integration (see above).

441

442 **Deletions and target site duplications**

443 We have analysed sizes of the deletions and target site duplications at the integration site by
444 examining insertions where both junctions of the insertion have been characterised by
445 amplicon sequencing. Deletions usually range between 1 and 50 bp, but larger deletions have
446 been observed as well, while target site duplications are shorter and less common. While the
447 simple SSA-like (single-strand annealing) repair model can explain deletions very well, target
448 site duplications are only possible when ‘sticky ends’ are created during creation of the DSBs.
449 The maximum size of a target site duplication is then the distance of the two staggered single-
450 strand breaks. Another explanation for target site duplications would be synthesis-dependent
451 strand annealing (SDSA) where a single strand 3'-end invades a homologous region (although
452 non-homologous regions are also possible) on the other end of the broken chromosome and
453 primes DNA-synthesis. This junction is pretty weak and might be aborted quickly. The repair
454 would then be completed by one the repair mechanisms available in *A. thaliana*.
455 Nevertheless, the distribution we observed indicates the creation of sticky ends with randomly
456 distributed length or blunt ends in most cases. Other random factors are the number of bases
457 which are degraded by exonucleases and the region where microhomology takes place
458 (determining the size of overhangig ends that are removed later). Microhomology-based base
459 pairing taking place at the very ends of the single stranded DNA of a sticky end during repair

460 leads to target site duplications, while microhomology-based base pairing distant from the
461 ends leads to removal of overhanging ends and therefore to deletions. Both of these factors
462 taken together (degradation of sticky ends and necessity of microhomology at the ends)
463 favour deletions, which fits to the observed insertion structures.

464 According to Salomon *et al.*(1998), the sizes of deletions observed during DSB repair range
465 from 2 bp to 1.2 kbp, which matches our observations. Because of our primer design (usually
466 around 450 bp from the predicted insertion site) we rarely found deletions larger than about
467 300 bp, but the deletion size distribution shows that the main fraction of deletions is up to 30
468 bp in length while larger deletions were less frequently observed with increasing size of the
469 deletion. The existence of an undetected high number of large deletions is therefore unlikely.

470 Nevertheless, due to FSTs present on both junctions of the insertion for a few GABI-Kat
471 lines, we could verify deletions of several kbp as shown in detail for the two cases in
472 GK_144F03 and GK_478B05, and also for a total of 12 cases where a gene not associated
473 with the FST has been deleted during integration (see Supplemental Information File SI8).

474 Furthermore we detected several inversions at the integration site. The occurrence of events
475 like large deletions and target site duplications should be considered during reverse genetics
476 which relies on T-DNA insertion mutants. It is important to always check both junctions of a
477 T-DNA insertion allele in order to exclude that phenotypes observed in the mutant are really
478 linked to the single gene apparent from an FST from just one border. With respect to the
479 mechanism of integration, inversions have also been observed during MMEJ
480 (microhomology-mediated end-joining) repair of DSB(McVey and Lee, 2008). This is
481 another piece of evidence that MMEJ plays a role for T-DNA integration. One explanation for
482 inversions is the occurrence of two DSB in close chromosomal neighbourhood. After repair,
483 the deleted fragment of genomic DNA gets incorporated right beside the T-DNA. It is
484 unlikely that inversions take place before T-DNA integration because all inversions we

485 detected were only found at one end of the T-DNA insertion, and not spanning the integration
486 site.

487

488 **Microhomology and fillers**

489 The analyses of microhomology and fillers indicates that microhomology seems to be a
490 prerequisite for integration most of the time, further supporting the NHEJ hypothesis in which
491 microhomology between 1 and 4 bp is expected (Gorbunova and Levy, 1999). Nevertheless,
492 we observed a significant number of cases with microhomology of up to 20 bp, indicating that
493 MMEJ might play a role more often than previously thought. The high abundance of fillers
494 and their origin highlights the importance of the SDSA pathway for T-DNA integration. The
495 fraction of insertion cases showing microhomology is significantly larger than what would be
496 expected as a random observation of sequence similarity at the junctions. We observed
497 microhomology of up to 20 bp at a regular basis. Even larger microhomology was not
498 observed, which can be explained by the lack of longer sequence similarity of the area
499 surrounding the T-DNA borders to sequences within the *A. thaliana* genome. Along with the
500 frequent presence of fillers these results support the hypothesis that T-DNA insertion is
501 largely mediated through the NHEJ pathway assisted by MMEJ. It would be an interesting
502 experiment to compare the sizes of microhomologies in integrations of mutants where Ku70,
503 Ku80 or XRCC4 are knocked out to check if MMEJ is used more often in such genetic
504 background.

505 Regarding the source of the sequences found in fillers, the majority was found to be derived
506 from the same chromosome. In addition, a large fraction of the source hits were found in very
507 close genomic proximity to the insertion site. This can be explained by assuming that the
508 migrating replication bubbles during SDSA find their target more likely in close genomic
509 neighbourhood. Fillers share microhomology to their adjacent T-DNA or genomic sequences
510 with half the rate of junctions without fillers. This fits to the assumption that a filler is created

511 at one of the open ends (genomic or T-DNA) using SDSA, without the use of microhomology
512 and that the other part is repaired using plant repair pathways utilizing microhomology. Thus,
513 only one side of the filler is expected to show microhomology, explaining the halved rate.
514 Fillers from templates that were not from the target genome have been observed rarely and
515 can be explained by the use of sequences as source of the filler that have been imported from
516 *Agrobacterium tumefaciens* to the side of integration.

517

518 **Processing of borders**

519 The analyses of the cut positions at the T-DNA border sequences confirmed the known
520 restriction site immediately upstream of the recognition sequence 5'-CAGGATATATT-
521 3'(Jasper et al., 1994) but also detected a large fraction of T-DNA border sequences that have
522 been further shortened. This varying degree of shortening might be depending on the
523 exposure time of the T-DNA to exonucleases until integration into a target DSB. The RB of
524 the T-strand is initially protected by VirD2 and therefore not exposed to exonuclease activity.
525 This is also reflected in our data by fewer basepair removals at the RB in the analysis of the
526 exact border cut. Nevertheless, there is a quite large number of integrations where some bases
527 are lost before or during integration.

528 During DSB repair by NHEJ or MMEJ, further deletions are possible by removal of
529 overhanging flaps not contributing to microhomology, which explains further shortened left
530 and right border sequences. Shortened border sequences can be observed in similar frequency
531 for both borders (although the size differs), which leads to the conclusion that the same DSB
532 repair pathway is used for processing of both borders during integration.

533 On both borders, extended fit of the border sequences by a few bases can be observed. This
534 short range extension is due to random sequence similarity of the genomic DNA to the border
535 sequence that has roughly a chance of 1 out of four to fit at the next position. In addition to
536 this "extended fit", some sequence of the binary vector neighbouring the T-DNA was found to

537 be integrated as well and fits to published reports for several transformed plants (Smith,
538 1998). Transfer of vector sequences has also been observed in yeast where sequences
539 following the border sequence can be transported into the host cell which is due to border
540 skipping or incomplete cutting of borders (Rolloos et al., 2014).

541

542 **Plant repair pathway for T-DNA integration**

543 Our results strongly support the hypothesis that integration fully utilises the plant repair
544 pathway for integration. The main DSB repair pathways in *A. thaliana* are NHEJ and MMEJ
545 supported by SDSA. The microhomologies, fillers and sizes of deletions or target site
546 duplications which we detected at insertion sites match the typical characteristics of DSB
547 repairs. A model for T-DNA integration incorporating these plant repair pathways is
548 summarised in Figure 7.

549 Just recently, two studies have been published (Mestiri et al., 2014; Park et al., 2015),
550 indicating that T-DNA integration still works when key proteins of the NHEJ pathway are
551 knocked out. Our assumption is that integration utilises available pathways, which might be
552 alternative repair pathways such as MMEJ, which play a larger role in the NHEJ deficient
553 lines. We would expect that in those lines the integration structures show longer stretches of
554 microhomology, a prediction that is supported by results from double-strand break repair
555 outcomes in Ku-deficient CHO cells where larger microhomology was observed (Feldmann et
556 al., 2000). We hypothesize that knockouts in genes required for MMEJ, such as the MRX-
557 complex consisting of Mre11, Rad50 and Xrs2(NBS1) (McVey and Lee, 2008), in addition to
558 the NHEJ knockouts might lead to lines unsusceptible for transformation.

559 The origin of the initial double strand breaks used for integration remains an open question.
560 Randomly occurring breaks might be a possible source but it seems unlikely that they are
561 occurring to an extent required for the random incorporation of T-DNA present somewhere in
562 the nucleus. This requires either a mechanism targeting the T-DNA to double-strand breaks

563 that are, for example, created by DNA topoisomerases during transcription or replication, or
564 the possibility to mediate their formation. However, such functionality has not yet been
565 assigned to one of the *Vir*-proteins associated with the T-DNA in plant cells.

566

567 **Methods**

568 **Origin and processing of confirmation sequences**

569 For confirmation of insertion alleles, confirmation amplicons were generated using a primer
570 specific for the T-DNA border from that the FST was derived and an insertion site or genome
571 specific primer (gsp) as described before (Huep et al., 2014). After Sanger sequencing with a
572 nested border specific primer and the gsp, the resulting sequences were processed using
573 PHRED and pregap4 from the staden package (Staden et al., 2000) for quality trimming.
574 Sequences for submission have been shortened to 1000 bp due to submission guidelines, the
575 shortening was done by leaving the T-DNA end of the sequence intact while removing bases
576 from the other end.

577

578 **Determination of the insertion position from confirmation sequences**

579 All analyses in this publication were based on the TAIR10 data version for the *A. thaliana*
580 genome. To derive a position for an insertion junction, a MEGABLAST vs. the TAIR10
581 genome sequence was performed. To focus on reliable sequence data we included the
582 following filters: (i) T-DNA sequence needed to be present at the 5' region (for sequences
583 sequenced with the T-DNA border primer) or at the 3' region (for sequences sequenced with
584 gsp) of the confirmation sequence; (ii) MEGABLAST hits in the genome required an e-value
585 of 1e-50 at most; if there were several sequences fulfilling these criteria for one junction of
586 the insertion, sequences sequenced with the T-DNA specific primer were preferred, and if
587 there were still multiple sequences the one with the lowest e-value was chosen. The position

588 of the start of the MEGABLAST hit was then determined as the insertion position of this
589 junction of the insertion.

590

591 **Determination of deletion sizes / target site duplication sizes**

592 To determine the structure of a given insertion site, we performed the confirmation PCR and
593 amplicon sequencing on both sites of the expected T-DNA, i.e. at the north and the south
594 junction of the T-DNA to the genome, even if there was only an FST for one of the junctions.
595 Primers were designed independently for each junction using the GABI-Kat primer design
596 tool (Huep et al., 2014). For insertions with unknown border identity at the 2nd junction,
597 which is the usual case for an insertion predicted from a single FST, we used LB as well as
598 RB specific T-DNA primers for the junction without FST. The insertion position for the
599 upstream and downstream junction of the T-DNA was computed as described above. The size
600 of the deletion was then calculated by the formula: (downstream_position -
601 upstream_position)-1. If this value is negative, north and south genomic junction sequences
602 are overlapping which indicates a target site duplication.

603

604 **Filtering of confirmation sequences for junction analysis**

605 Confirmation sequences were filtered to make sure that only one and the highest quality
606 sequence was evaluated for a given junction. Only the outermost border (BLAST hit) was
607 evaluated, and we required that this outermost border was present in the correct orientation.
608 One confirmation sequence for each junction of an insertion site was chosen according to the
609 following criteria: (i) presence of T-DNA and *A. thaliana* genomic sequence was required; (ii)
610 sequences sequenced with the gsp were preferred because they usually contain a larger part of
611 the T-DNA sequence since the read often reaches into the sequence contributed by the border
612 primer; (iii) if there were still multiple sequences available the one with the highest e-value
613 for the hit in the *A. thaliana* genome was preferred. Of all 30,507 confirmation sequences,

614 11,802 were used for further analysis. Out of these, 699 originated from the RB, 11,103 from
615 the LB, and each single read represented one unique T-DNA/genome junction. For SALK
616 lines, 880 sequences selected from 2,469 confirmation sequences were analysed, 852 for the
617 LB and 28 for the RB.

618

619 **Sequence analyses of T-DNA borders**

620 The limits of T-DNA sequences present at a given insertion site, possibly along with
621 sequences derived from the adjacent plasmid vector, were determined based on BLAST
622 analyses of the sequences that were filtered as described above. First, the BLAST analyses
623 were performed vs. the complete binary vector sequence, and second against the T-DNA
624 sequence, both with an e-value cutoff of 10 and no filtering for low complexity regions. The
625 results were analysed using a dedicated Perl-script along with the LIMS database used at
626 GABI-Kat (Kleinboelting et al., 2012) for storage of sequences and BLAST results. The script
627 compared the two relative positions of the two hits. If the BLAST hit against the vector was
628 longer than the hit against the T-DNA part, a part of the vector was integrated as well. This
629 was indicated by a positive output value from the script. Zero indicated an exact cut of the T-
630 DNA right at the outer end of the sequences defined as LB or RB (see Figure 1). A negative
631 output value indicated that parts of the border sequences were missing.

632

633 **Determination of microhomology at integration sites**

634 In order to determine microhomology or fillers at a given integration site, additional BLAST
635 analyses (with the filtered sequences and with respect to the previously performed ones, see
636 paragraph above) vs. the *A. thaliana* TAIRv10 genome sequence were performed with an e-
637 value cutoff of 5e-3 and no filtering for low complexity regions. If the BLAST hit vs. the
638 plant genome follows directly before/after the hit vs. the vector, integration occurred exactly
639 without use of microhomology and without creating fillers, giving zero as result. If BLAST

640 hits are overlapping, microhomology has been detected and is assigned a negative value in the
641 amount of overlapping base pairs. A positive value is assigned, if there is a gap between the
642 vector hit and the plant genomic hit, indicating a filler. Although these junctions are classified
643 as junctions with fillers (and without microhomology) in this analysis, it is possible that there
644 is microhomology between the filler and the T-DNA or plant genomic DNA. In order to
645 identify microhomology of fillers, the origin of the fillers had to be determined first.

646

647 **Examination of fillers**

648 All fillers of at least 15 bp (see above) were extracted from the respective confirmation
649 sequence with an overlap of 20 bp to the T-DNA part and the genomic part of the
650 confirmation sequences and stored separately (see Supplemental Information File SI9 for GK
651 and SI10 for SALK data). Shorter sequences were not considered because they would not
652 allow generating meaningful BLAST results. BLAST runs vs. sequence data sets from *A.*
653 *thaliana*, *A. tumefaciens*, chloroplast, mitochondria, T-DNA from LB to RB and vector
654 backbone (without the T-DNA) were performed with an e-value cutoff of 10.0 in order to
655 identify the origin of those fillers. To exclude hits mainly based on the overlap of 20 bp, hits
656 covering the main part of the original filler were filtered in a first step. If these criteria did not
657 result in a unique hit, the one with the best score among them was chosen. In about 22 % of
658 cases several "best" BLAST hits were obtained, covering the same length of filler-query with
659 identical scores. Initially, we excluded these cases. As a result, 1,162 sequences were left (389
660 hits on the same chromosome, 756 on other chromosomes, 15 on the *A. tumefaciens* genome
661 and 2 on the mitochondrial genome). Hits within the T-DNA or Vector were filtered by
662 excluding these ambiguous cases, mostly due to sequence identity of a part of the T-DNA
663 with the vector backbone (the T-DNA allows plasmid rescue). We decided to assign
664 ambiguous cases to the most probable origin in the order of the ranking observed for almost
665 80 % of all cases. First, hit on the same chromosome, if in doubt the one closest to the

666 insertion position; second, hit on a chromosome other than the one that contained the studied
667 filler; third, hit within the T-DNA from LB to RB; fourth, hit on the vector backbone; fifth, hit
668 on the *Agrobacterium tumefaciens* genome (circular, linear, plasmid); sixth, hit on chloroplast
669 or mitochondrial DNA sequence. That measure did not change the overall trend of the results,
670 but avoided loss of hits on the T-DNA or vector backbone. If the hit was on the same
671 chromosome, it was used to compute the distance of the source position of the filler to the
672 insertion position. Microhomology or filler between the filler and the neighbouring sequences
673 was then determined by evaluating the range of the BLAST hit – if it reaches into the 20 bp
674 overlap, microhomology is present.

675

676 **Sequencing of line GK_144F03**

677 Genomic DNA (820 ng) prepared from a pool of 50 T3 offspring seedlings homozygous for
678 the insertion allele affecting the T-DNA insertion locus close to *At3g48060* was used for
679 preparing paired end libraries according to the Illumina TruSeqv2 Sample Prep Kit.

680 Homozygosity was proven for the parental T2 plant as well as for the pooled T3 plants using
681 primer combinations as described by (Bolle et al., 2013). Sequencing was performed on an
682 Illumina MiSeq sequencing instrument with 2×250 nt paired-end reads using the MiSeqv2
683 chemistry. Reads were trimmed by Trimmomatic version 0.27 (Bolger et al., 2014) and
684 mapped to the *A. thaliana* TAIRv10 genome using BWA (Li and Durbin, 2009) with default
685 parameters. The data were visualised using Genomeview (Abeel et al., 2012).

686

687 **Histochemical assays and microscopy for tt mutant analyses**

688 Microscopy, binocular analysis, image acquisition as well as DPBA staining, DMACA
689 staining and norflurazon treatment of plant material was performed as described previously
690 (Appelhagen et al., 2014).

691

692 **Accession numbers and links to external databases**

693 All confirmation sequences have been submitted to ENA/GenBank, the accession numbers
694 are LN484267- LN515397. Illumina read data from sequencing of GK_144F03 genomic
695 DNA were deposited at the NCBI Short Read Archive (SRA;
696 <http://www.ncbi.nlm.nih.gov/sra/>) under the accession number SRX879613.

697

698

699 **Author Contributions**

700 NK, GH and BW conceived the research approach, interpreted the data and wrote the
701 manuscript. NK, GH, YL and BW analysed the data. GH conducted and supervised the wet-
702 lab experiments. PV designed and performed all Sanger and NGS sequencing. NK carried out
703 database programming and bioinformatic analyses. BW designed the research outline. IA
704 identified and characterised the GK_478B05/*tt6* mutant and contributed to writing the
705 manuscript. All authors read and approved the manuscript.

706

707

708 **Acknowledgements**

709 The authors thank Mario Rosso, the MPI for Plant Breeding Research and all former co-
710 workers in the project for their contribution to GABI-Kat, Ute Buerstenbinder, Renate Harder,
711 Eliane Quittschau, Helene Schellenberg, Nina Schmidt, Andrea Voigt, Marja-Leena Wilke for
712 technical assistance, and Malte Mattheis for database programming to identify inversions. The
713 work described in this article was funded by the German Federal Ministry of Education and
714 Research (BMBF) in the context of the German plant genomics program GABI
715 (Förderkennzeichen 0313855). We acknowledge the financial support of the German
716 Research Foundation (DFG) and the Open Access Publication Fund of Bielefeld University
717 for the article processing charge.

718

719 **Competing interests**

720 The authors declare that they have no competing interests.

721

722

723 **References**

724

- 725 Abeel, T., Van Parys, T., Saeys, Y., Galagan, J., and Van de Peer, Y. (2012). GenomeView: a
726 next-generation genome browser. *Nucleic Acids Res* 40.
- 727 Alonso, J.M., Stepanova, A.N., Lisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K.,
728 Zimmerman, J., Barajas, P., Cheuk, R., et al. (2003). Genome-wide insertional
729 mutagenesis of *Arabidopsis thaliana*. *Science* 301:653-657.
- 730 Anand, A., Vaghchhipawala, Z., Ryu, C.M., Kang, L., Wang, K., del-Pozo, O., Martin, G.B.,
731 and Mysore, K.S. (2007). Identification and characterization of plant genes involved in
732 Agrobacterium-mediated plant transformation by virus-induced gene silencing. *Mol*
733 *Plant Microbe Interact* 20:41-52.
- 734 Appelhagen, I., Thiedig, K., Nordholt, N., Schmidt, N., Huet, G., Sagasser, M., and
735 Weisshaar, B. (2014). Update on *transparent testa* mutants from *Arabidopsis thaliana*:
736 characterisation of new alleles from an isogenic collection. *Planta* 240:955-970.
- 737 Bolger, M.E., Weisshaar, B., Scholz, U., Stein, N., Usadel, B., and Mayer, K.F. (2014). Plant
738 genome sequencing - applications for crop improvement. *Curr Opin Biotechnol* 26:31-
739 37.
- 740 Bolle, C., Huet, G., Kleinbolting, N., Haberer, G., Mayer, K., Leister, D., and Weisshaar, B.
741 (2013). GABI-DUPLO: a collection of double mutants to overcome genetic
742 redundancy in *Arabidopsis thaliana*. *Plant J* 75:157-171.
- 743 Britt, A.B., and May, G.D. (2003). Re-engineering plant gene targeting. *Trends Plant Sci*
744 8:90-95.
- 745 Brunaud, V., Balzergue, S., Dubreucq, B., Aubourg, S., Samson, F., Chauvin, S., Bechtold,
746 N., Cruaud, C., DeRose, R., Pelletier, G., et al. (2002). T-DNA integration into the
747 *Arabidopsis* genome depends on sequences of pre-insertion sites. *EMBO Rep* 3:1152-
748 1157.
- 749 Chilton, M.D., and Que, Q. (2003). Targeted integration of T-DNA into the tobacco genome
750 at double-stranded breaks: new insights on the mechanism of T-DNA integration.
751 *Plant Physiol* 133:956-965.
- 752 Citovsky, V., Wong, M.L., and Zambryski, P. (1989). Cooperative interaction of
753 Agrobacterium VirE2 protein with single-stranded DNA: implications for the T-DNA
754 transfer process. *Proceedings of the National Academy of Science of the United States*
755 *of America* 86:1193-1197.
- 756 De Buck, S., Jacobs, A., Van Montagu, M., and Depicker, A. (1999). The DNA sequences of
757 T-DNA junctions suggest that complex T-DNA loci are formed by a recombination
758 process resembling T-DNA integration. *Plant J* 20:295-304.
- 759 De Neve, M., De Buck, S., Jacobs, A., Van Montagu, M., and Depicker, A. (1997). T-DNA
760 integration patterns in co-transformed plant cells suggest that T- DNA repeats
761 originate from co-integration of separate T-DNAs. *Plant J* 11:15-29.
- 762 Dürrenberger, F., Cramer, A., Hohn, B., and Koukolíková-Nicola, Z. (1989). Covalently
763 bound VirD2 protein of *Agrobacterium tumefaciens* protects the T-DNA from
764 exonucleolytic degradation. *Proceedings of the National Academy of Science of the*
765 *United States of America* 86:9154-9158.
- 766 Feldmann, E., Schmiemann, V., Goedecke, W., Reichenberger, S., and Pfeiffer, P. (2000).
767 DNA double-strand break repair in cell-free extracts from Ku80-deficient cells:
768 implications for Ku serving as an alignment factor in non-homologous DNA end
769 joining. *Nucleic Acids Res* 28:2585-2596.
- 770 Forsbach, A., Schubert, D., Lechtenberg, B., Gils, M., and Schmidt, R. (2003). A
771 comprehensive characterization of single-copy T-DNA insertions in the *Arabidopsis*
772 *thaliana* genome. *Plant Mol Biol* 52:161-176.

773 Gelvin, S.B. (1998). The introduction and expression of transgenes in plants. *Curr Opin*
774 *Biotechnol* 9:227-232.

775 Gelvin, S.B. (2010). Finding a way to the nucleus. *Current Opinion in Microbiology* 13:53-
776 58.

777 Gheysen, G., Villarroel, R., and Van Montagu, M. (1991). Illegitimate recombination in
778 plants: a model for T-DNA integration. *Genes Dev* 5:287-297.

779 Gorbunova, V.V., and Levy, A.A. (1999). How plants make ends meet: DNA double-strand
780 break repair. *Trends Plant Sci* 4:263-269.

781 Hoekema, A., Hirsch, P.R., Hooykaas, P.J.J., and Schilperoort, R.A. (1983). A binary plant
782 vector strategy based on separation of vir- and T-region of the *Agrobacterium*
783 *tumefaciens* Ti-plasmid. *Nature* 303:179-180.

784 Huep, G., Kleinboelting, N., and Weisshaar, B. (2014). An easy-to-use primer design tool to
785 address paralogous loci and T-DNA insertion sites in the genome of *Arabidopsis*
786 *thaliana*. *Plant Methods* 10.

787 Jasper, F., Koncz, C., Schell, J., and Steinbiss, H.H. (1994). *Agrobacterium* T-strand
788 production in vitro: sequence-specific cleavage and 5' protection of single-stranded
789 DNA templates by purified VirD2 protein. *Proc Natl Acad Sci USA* 91:694-698.

790 Kim, S., Veena, and Gelvin, S. (2007). Genome-wide analysis of *Agrobacterium* T-DNA
791 integration sites in the *Arabidopsis* genome generated under non-selective conditions.
792 *Plant J* 51:779-791.

793 Kirik, A., Salomon, S., and Puchta, H. (2000). Species-specific double-strand break repair and
794 genome evolution in plants. *EMBO J* 19:5562-5566.

795 Kleinboelting, N., Huep, G., Kloetgen, A., Viehoveer, P., and Weisshaar, B. (2012). GABI-
796 Kat SimpleSearch: new features of the *Arabidopsis thaliana* T-DNA mutant database.
797 *Nucleic Acids Res* 40:D1211-D1215.

798 Krizkova, L., and Hroudá, M. (1998). Direct repeats of T-DNA integrated in tobacco
799 chromosome: characterization of junction regions. *Plant J* 16:673-680.

800 Lacroix, B., Vaidya, M., Tzfira, T., and Citovsky, V. (2005). The VirE3 protein of
801 *Agrobacterium* mimics a host cell function required for plant genetic transformation.
802 *EMBO J* 24:428-437.

803 Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler
804 transform. *Bioinformatics* 25:1754-1760.

805 Li, Y., Rosso, M.G., Ulker, B., and Weisshaar, B. (2006). Analysis of T-DNA insertion site
806 distribution patterns in *Arabidopsis thaliana* reveals special features of genes without
807 insertions. *Genomics* 87:645-652.

808 Liang, Z., and Tzfira, T. (2013). In vivo formation of double-stranded T-DNA molecules by
809 T-strand priming. *Nature Communications* 4.

810 Magori, S., and Citovsky, V. (2011). *Agrobacterium* counteracts host-induced degradation of
811 its effector F-box protein. *Science Signaling* 4.

812 Mayerhofer, R., Koncz-Kalman, Z., Nawrath, C., Bakkeren, G., Cramer, A., Angelis, K.,
813 Redei, G.P., Schell, J., Hohn, B., and Koncz, C. (1991). T-DNA integration: a mode of
814 illegitimate recombination in plants. *EMBO J* 10:697-704.

815 McVey, M., and Lee, S.E. (2008). MMEJ repair of double-strand breaks (director's cut):
816 deleted sequences and alternative endings. *Trends Genet* 24:529-538.

817 Mestiri, I., Norre, F., Gallego, M.E., and White, C.I. (2014). Multiple host-cell recombination
818 pathways act in *Agrobacterium*-mediated transformation of plant cells. *Plant J* 77.

819 Mysore, K.S., Kumar, C.T., and Gelvin, S.B. (2000a). *Arabidopsis* ecotypes and mutants that
820 are recalcitrant to *Agrobacterium* root transformation are susceptible to germ-line
821 transformation. *Plant J* 21:9-16.

822 Mysore, K.S., Nam, J., and Gelvin, S.B. (2000b). An Arabidopsis histone H2A mutant is
823 deficient in Agrobacterium T-DNA integration. *Proceedings of the National Academy*
824 *of Science of the United States of America* 97:948-953.

825 Nam, J., Matthysse, A.G., and Gelvin, S.B. (1997). Differences in susceptibility of
826 Arabidopsis ecotypes to crown gall disease may result from a deficiency in T-DNA
827 integration. *Plant Cell* 9:317-333.

828 Nam, J., Mysore, K.S., and Gelvin, S.B. (1998). Agrobacterium tumefaciens transformation
829 of the radiation hypersensitive Arabidopsis thaliana mutants uvh1 and rad5. *Mol Plant*
830 *Microbe Interact* 11:1136-1141.

831 O'Malley, R.C., and Ecker, J.R. (2010). Linking genotype to phenotype using the Arabidopsis
832 unimutant collection. *Plant J* 61:928-940.

833 Orel, N., and Puchta, H. (2003). Differences in the processing of DNA ends in Arabidopsis
834 thaliana and tobacco: possible implications for genome evolution. *Plant Mol Biol*
835 51:523-531.

836 Park, S.Y., Vaghchhipawala, Z., Vasudevan, B., Lee, L.Y., Shen, Y., Singer, K., Waterworth,
837 W.M., Zhang, Z.J., West, C.E., Mysore, K.S., et al. (2015). Agrobacterium T-DNA
838 integration into the plant genome can occur without the activity of key non-
839 homologous end-joining proteins. *Plant J*.

840 Peralta, E.G., and Ream, L.W. (1985). T-DNA border sequences required for crown gall
841 tumorigenesis. *Proc Natl Acad Sci USA* 82:5112-5116.

842 Pitzschke, A., and Hirt, H. (2010). New insights into an old story: Agrobacterium-induced
843 tumour formation in plants by plant transformation. *EMBO Journal* 29:1021-1032.

844 Que, Q., Elumalai, S., Li, X., Zhong, H., Nalapalli, S., Schweiner, M., Fei, X., Nuccio, M.,
845 Kelliher, T., Gu, W., et al. (2014). Maize transformation technology development for
846 commercial event generation. *Frontiers in Plant Science* 5.

847 Ray, A., and Langer, M. (2002). Homologous recombination: ends as the means. *Trends Plant*
848 *Sci* 7:435-440.

849 Rodenburg, K.W., de Groot, M.J., Schilperoort, R.A., and Hooykaas, P.J. (1989). Single-
850 stranded DNA used as an efficient new vehicle for transformation of plant protoplasts.
851 *Plant Mol Biol* 13:711-719.

852 Rolloos, M., Dohmen, M.H., Hooykaas, P.J., and van der Zaal, B.J. (2014). Involvement of
853 Rad52 in T-DNA circle formation during Agrobacterium tumefaciens-mediated
854 transformation of *Saccharomyces cerevisiae*. *Molecular Microbiology* 91:1240-1251.

855 Salomon, S., and Puchta, H. (1998). Capture of genomic and T-DNA sequences during
856 double-strand break repair in somatic plant cells. *EMBO J* 17:6086-6095.

857 Schell, J., Van Montagu, M., De Beuckeleer, M., De Block, M., Depicker, A., De Wilde, M.,
858 Engler, G., Genetello, C., Hernalsteens, J.P., Holsters, M., et al. (1979). Interactions
859 and DNA transfer between Agrobacterium tumefaciens, the Ti-plasmid and the plant
860 host. *Proceedings of the Royal Society of London* 204:251-266.

861 Scholl, R.L., May, S.T., and Ware, D.H. (2000). Seed and molecular resources for
862 Arabidopsis. *Plant Physiol* 124:1477-1480.

863 Singer, K., Shibolet, Y.M., Li, J., and Tzfira, T. (2012). Formation of complex
864 extrachromosomal T-DNA structures in Agrobacterium tumefaciens-infected plants.
865 *Plant Physiol* 160:511-522.

866 Smith, N. (1998). More T-DNA than meets the eye. *Trends Plant Sci* 3:85.

867 Staden, R., Beal, K.F., and Bonfield, J.K. (2000). The Staden package, 1998. *Methods Mol*
868 *Biol* 132:115-130.

869 Strizhov, N., Li, Y., Rosso, M.G., Viehoveer, P., Dekker, K.A., and Weisshaar, B. (2003).
870 High-throughput generation of sequence indexes from T-DNA mutagenized
871 Arabidopsis thaliana lines. *BioTechniques* 35:1164-1168.

- 872 Szabados, L., Kovacs, I., Oberschall, A., Abraham, E., Kerekes, I., Zsigmond, L., Nagy, R.,
873 Alvarado, M., Krasovskaja, I., Gal, M., et al. (2002). Distribution of 1000 sequenced
874 T-DNA tags in the Arabidopsis genome. *Plant J* 32:233-242.
- 875 Thole, V., Alves, S.C., Worland, B., Bevan, M.W., and Vain, P. (2009). A protocol for
876 efficiently retrieving and characterizing flanking sequence tags (FSTs) in
877 *Brachypodium distachyon* T-DNA insertional mutants. *Nat Protoc* 4:650-661.
- 878 Tinland, B., Hohn, B., and Puchta, H. (1994). *Agrobacterium tumefaciens* transfers single-
879 stranded transferred DNA (T-DNA) into the plant cell nucleus. *Proc Natl Acad Sci*
880 *USA* 91:8000-8004.
- 881 Tzfira, T., Frankman, L.R., Vaidya, M., and Citovsky, V. (2003). Site-specific integration of
882 *Agrobacterium tumefaciens* T-DNA via double-stranded intermediates. *Plant Physiol*
883 133:1011-1023.
- 884 Tzfira, T., Li, J., Lacroix, B., and Citovsky, V. (2004). *Agrobacterium* T-DNA integration:
885 molecules and models. *Trends Genet* 20:375-383.
- 886 Ulker, B., Li, Y., Rosso, M.G., Logemann, E., Somssich, I.E., and Weisshaar, B. (2008). T-
887 DNA-mediated transfer of *Agrobacterium tumefaciens* chromosomal DNA into plants.
888 *Nat Biotechnol* 26:1015-1017.
- 889 Vaghchhipawala, Z.E., Vasudevan, B., Lee, S., Morsy, M.R., and Mysore, K.S. (2012).
890 *Agrobacterium* may delay plant nonhomologous end-joining DNA repair via XRCC4
891 to favor T-DNA integration. *Plant Cell* 24:4110-4123.
- 892 van Attikum, H., Bundock, P., Overmeer, R.M., Lee, L.Y., Gelvin, S.B., and Hooykaas, P.J.
893 (2003). The Arabidopsis AtLIG4 gene is required for the repair of DNA damage, but
894 not for the integration of *Agrobacterium* T-DNA. *Nucleic Acids Res* 31:4247-4255.
- 895 Van Larebeke, N., Engler, G., Holsters, M., Van den Elsacker, S., Zaenen, I., Schilperoort,
896 R.A., and Schell, J. (1974). Large plasmid in *Agrobacterium tumefaciens* essential for
897 crown gall-inducing ability. *Nature* 252:169-170.
- 898 Wang, K., Herrera-Estrella, L., Van Montagu, M., and Zambryski, P. (1984). Right 25 bp
899 terminus sequence of the nopaline T-DNA is essential for and determines direction of
900 DNA transfer from *agrobacterium* to the plant genome. *Cell* 38:455-462.
- 901 Weaver, D.T. (1995). What to do at an end: DNA double-strand-break repair. *Trends Genet*
902 11:388-392.
- 903 Windels, P., De Buck, S., Van Bockstaele, E., De Loose, M., and Depicker, A. (2003). T-
904 DNA integration in Arabidopsis chromosomes. Presence and origin of filler DNA
905 sequences. *Plant Physiol* 133:2061-2068.
- 906 Yoshiyama, K.O., Sakaguchi, K., and Kimura, S. (2013). DNA Damage Response in Plants:
907 Conserved and Variable Response Compared to Animals. *Biology* 2:1338-1356.
- 908 Zambryski, P., Holsters, M., Kruger, K., Depicker, A., Schell, J., Van Montagu, M., and
909 Goodman, H. (1980). Tumor DNA structure in plant cells transformed by *A.*
910 *tumefaciens*. *Science* 209:1385-1391.

911
912
913

914 **Figure legends**

915 **Figure 1: Distribution of cut positions in integrated T-DNA border sequences**

916 Locations of cuts in the border sequence or T-DNA are shown on the x-axis. Zero (0)
917 corresponds to the end of the T-DNA defined by the LB (A) or RB (B) sequence, only T-
918 DNA sequences close to LB and RB are shown, the main part is represented by //. Negative
919 values mean further shortened borders. Positive values correspond to insertion of sequences
920 beyond the border displaying continuous sequence similarity to the T-DNA vector backbone.
921 The highest peak is at the expected locations (-3 for LB and -22 for RB), further shortened
922 border sequences are quite common.

923

924 **Figure 2: Size distribution of microhomologies and fillers at integration sites**

925 Length of microhomologous sequences and sequences present between the end of T-DNA and
926 plant genomic sequences at the insertion sites (see text for definition). A positive value
927 corresponds to a filler between T-DNA and plant genomic DNA, a negative value to a
928 microhomology of the size indicated in bp. Zero means a seamless transition from T-DNA to
929 plant genomic DNA sequence. Cases from RB and LB plotted together show a significant
930 portion of cases with microhomology as well as the common occurrence of fillers (A).
931 Microhomology and fillers can be found at RB and LB in similar size distributions (B, C).

932

933 **Figure 3: Characteristics of fillers**

934 Most of the fillers originate from the same chromosome with a bias towards close genomic
935 neighbourhood to the site of insertion (A,B,C). There was no detectable bias for a specific
936 chromosome as a source of fillers originating from another chromosome, neither when all
937 fillers were evaluated nor when the fillers from insertions of a single chromosome were
938 studied (D). Microhomology of a few bases as well as further fillers can be observed between
939 the filler sequence and the neighbouring *A. thaliana* genomic DNA (E) or T-DNA (F).

940

941 **Figure 4: Size distribution of deletions / target site duplications**

942 Length of deletions and target site duplications detected at insertion sites. These changes of
943 the pre-insertion site sequences occur on a regular basis with a bias towards smaller deletions.
944 Deletions larger than 100 bp as well as target site duplications larger than 20 bp do exist but
945 were observed rarely (see text for details).

946

947 **Figure 5: Coverage plot for the large deletion in line GK_144F03**

948 Display of a region on chromosome 3 that is affected by the deletion caused by the T-DNA
949 insertion. The cumulative coverage for both strands is shown and indicates a deletion
950 affecting three genes of which *At3g48070* is lost completely. The mapping data were
951 visualised using Genomeview (Abeel et al., 2012).

952

953 **Figure 6: Large deletion in line GK_478B05 affecting *At3g51240* and causing the**
954 **corresponding *tt6* phenotype**

955 GK_478B05.04 which is homozygous for the insertion shows the same phenotype as the *tt6*
956 knockout GK_292E08. GK_478B05.08 (heterozygous sibling) shows the WT-phenotype. The
957 phenotypical characterization confirms the removal of the gene *At3g51240*. Flavonol
958 accumulation in seedlings was visualised with Diphenyl boric acid 2-aminoethylester (DPBA)
959 and detected using epifluorescence microscopy in homozygous (.04 T3) and heterozygous
960 (.08 T3) plants of GK_478B05 and wild-type as well as *tt6-2* and *tt7-7* plants as controls.
961 Anthocyanins in seedlings were detected using microscopy after bleaching with norflurazon.
962 The proanthocyanidin content of mature seeds was visualised with DMACA staining and
963 detected with a binocular.

964

965 **Figure 7: Repair mechanisms used for integration**

966 Summary of double-strand breaks repair mechanisms hypothesized to contribute to T-DNA
967 integration. DSBs occurring for various reasons within the host genome are targets for
968 integration of a double-stranded T-DNA molecule. Integration occurs mainly via the NHEJ
969 and MMEJ pathways, both requiring some microhomology. Accurate repair via simple
970 ligation or HR is very rare and their characteristics almost absent from our data. Before repair
971 takes place, SDSA may create fillers of additional DNA sequence before one of the repair
972 pathways leads to an intact double stranded DNA. This may also explain the presence of
973 inversions at the integration site, when two double-strand breaks occur within close genomic
974 proximity and the fragment gets reincorporated in opposite orientation.

975

976 **Tables**

977 **Table 1: Source of fillers. A total of 1,508 fillers of different length were subjected to**
 978 **BLAST analysis, and results classified according to origin of the hit sequence.**

BLAST target	Number of hits
<i>A. thaliana</i> , same chromosome	548 / 1,508 (36.3 %)
<i>A. thaliana</i> , other chromosome	883 / 1,508 (58.6 %, 14.7 % per chromosome)
GABI-Kat T-DNA within 100 bp distance to LB/RB	46 / 1,508 (3.1 %)
GABI-Kat T-DNA outside 100 bp distance to LB/RB	2 / 1,508 (0.1 %)
GABI-Kat vector backbone within 100 bp distance to LB/RB	1 / 1,508 (0.1 %)
GABI-Kat vector backbone outside 100 bp distance to LB/RB	8 / 1,508 (0.5 %)
<i>A. tumefaciens</i> (linear, circular, plasmid)	16 / 1,508 (1.1 %)
chloroplast genome	0 / 1,508 (0 %)
mitochondrial genome	2 / 1,508 (0.1 %)
unknown	2 / 1,508 (0.1%)

979

980

981 **Table 2: Inversions in GABI-Kat.**

Line	BAC	next gene (AGI)	Position of inversion	Size of inversion
GK_091G10	F22I13	At4g38380	Chr4:17975281-17975444	164 bp
GK_041A02	F24M12	At3g51057	Chr3:18963229-18963303	75 bp
GK_526B04	T17B22	At3g03120	Chr3:717349-717511	163 bp
GK_118F06	T2H3	At4g02310	Chr4:1014928-1014979	52 bp
GK_692B09	F4D11	At4g32640	Chr4:15742746-15742807	62 bp
GK_182H05	F13I12	At3g47110	Chr3:17348070-17348138	69 bp
GK_135D02	T32G6	At2g41570	Chr2:17338757-17338972	216 bp
GK_545H03	MEE6	At5g41050	Chr5:16434876-16434954	79 bp
GK_401E04	F6E21	At4g31170	Chr4:15155129-15155593	465 bp
GK_072H03	T28M21	At2g40030	Chr2:16716822-16717129	308 bp

982

983

984 **List of Supplemental Information Files**

985 **Supplemental Information File SI1:** Tab-separated file containing the results for border cut
986 and microhomology / filler analysis of all evaluated confirmation sequences.

987 **Supplemental Information File SI2:** Tab-separated file containing lines with fillers that
988 originate from *A. tumefaciens*.

989 **Supplemental Information File SI3:** Tab-separated file containing the results of the detailed
990 filler analysis (origin of filler and microhomology to neighbouring sequences).

991 **Supplemental Information File SI4:** List of deletions and target site duplications at the
992 integration site detected in the GK data.

993 **Supplemental Information File SI5:** Tab-separated file containing the results for border cut
994 and microhomology / filler analysis of all evaluated confirmation sequences for SALK lines.

995 **Supplemental Information File SI6:** Tab-separated file containing the results of the detailed
996 filler analysis (origin of filler and microhomology to neighbouring sequences) for SALK
997 lines.

998 **Supplemental Information File SI7:** Tab-separated file with a list of deletions and target site
999 duplications at the integration site detected in the SALK data.

1000 **Supplemental Information File SI8:** Tab-separated file containing a list of genes affected by
1001 larger deletions.

1002 **Supplemental Information File SI9:** FASTA-file containing the filler sequences subjected
1003 for further analysis.

1004 **Supplemental Information File SI10:** FASTA-file containing the filler sequences from
1005 SALK data subjected for further analysis.

1006

1007 **List of abbreviations used**

1008 DSB: Double-strand break

1009 DSBR: Double-strand break repair

1010 gsp: gene specific primer

1011 HR: Homologous recombination

1012 LB: left border

1013 MMEJ: Microhomology-mediated end joining

1014 NHEJ: Non-homologous end joining

1015 RB: right border

1016 SDSA: Synthesis-dependent strand annealing

1017 SSA: Single strand annealing

1018

1019

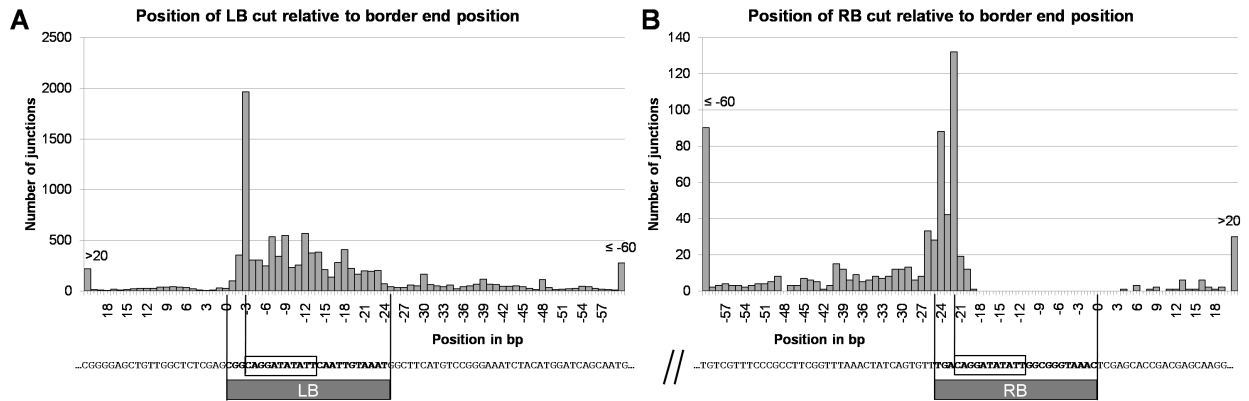


Figure 1: Distribution of cut positions in integrated T-DNA border sequences

Locations of cuts in the border sequence or T-DNA are shown on the x-axis. Zero (0) corresponds to the end of the T-DNA defined by the LB (A) or RB (B) sequence, only T-DNA sequences close to LB and RB are shown, the main part is represented by //. Negative values mean further shortened borders. Positive values correspond to insertion of sequences beyond the border displaying continuous sequence similarity to the T-DNA vector backbone. The highest peak is at the expected locations (-3 for LB and -22 for RB), further shortened border sequences are quite common.

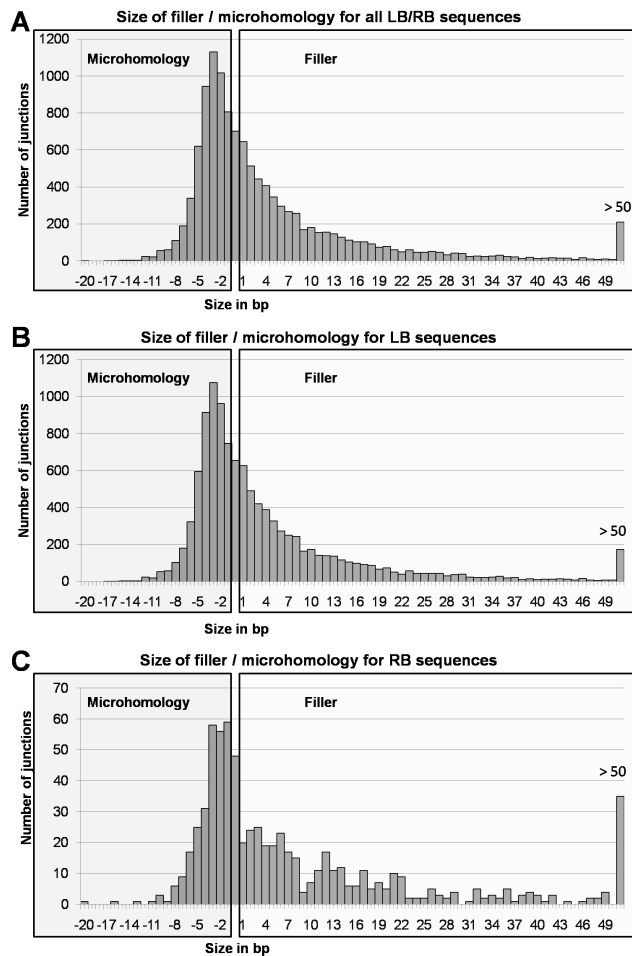


Figure 2: Size distribution of microhomologies and fillers at integration sites

Length of microhomologous sequences and sequences present between the end of T-DNA and plant genomic sequences at the insertion sites (see text for definition). A positive value corresponds to a filler between T-DNA and plant genomic DNA, a negative value to a microhomology of the size indicated in bp. Zero means a seamless transition from T-DNA to plant genomic DNA sequence. Cases from RB and LB plotted together show a significant portion of cases with microhomology as well as the common occurrence of fillers (A). Microhomology and fillers can be found at RB and LB in similar size distributions (B, C).

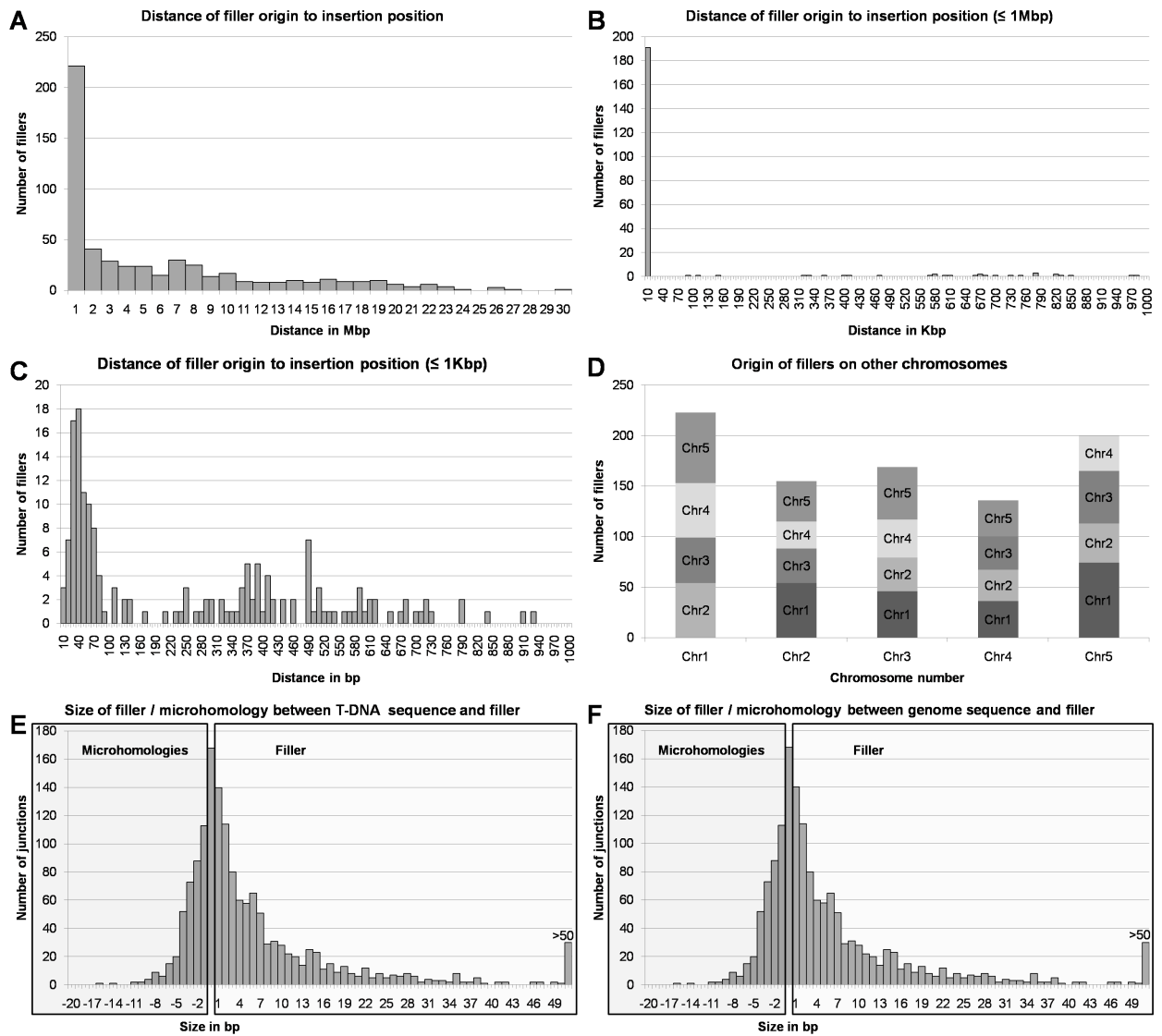


Figure 3: Characteristics of fillers

Most of the fillers originate from the same chromosome with a bias towards close genomic neighbourhood to the site of insertion (A,B,C). There was no detectable bias for a specific chromosome as a source of fillers originating from another chromosome, neither when all fillers were evaluated nor when the fillers from insertions of a single chromosome were studied (D). Microhomology of a few bases as well as further fillers can be observed between the filler sequence and the neighbouring *A. thaliana* genomic DNA (E) or T-DNA (F).

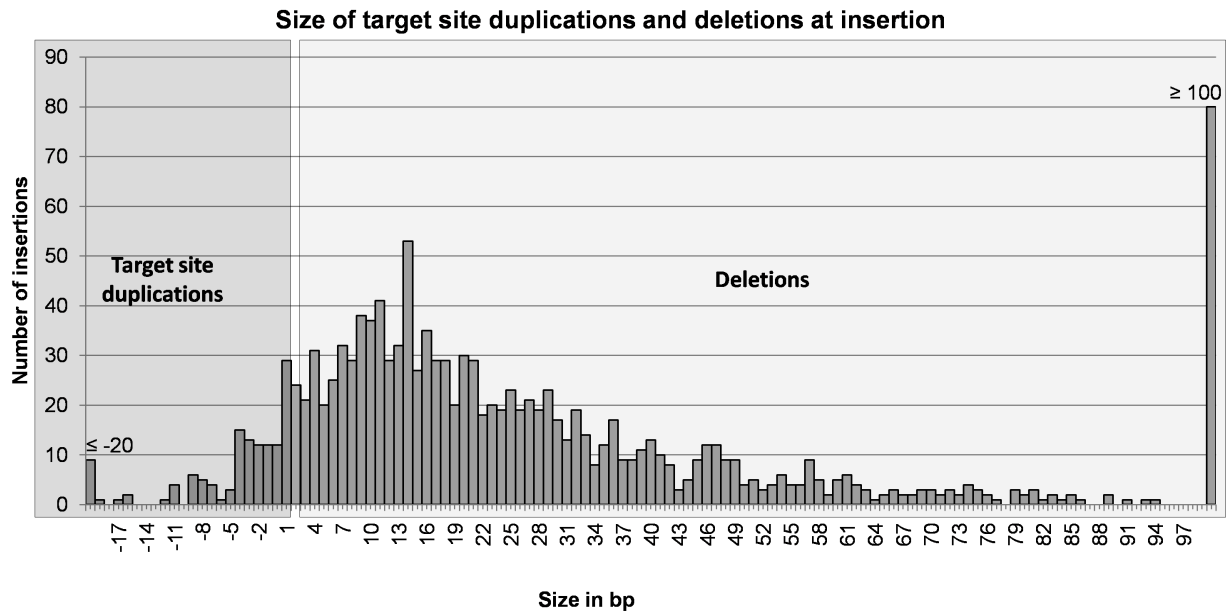


Figure 4: Size distribution of deletions / target site duplications

Length of deletions and target site duplications detected at insertion sites. These changes of the pre-insertion site sequences occur on a regular basis with a bias towards smaller deletions. Deletions larger than 100 bp as well as target site duplications larger than 20 bp do exist but were observed rarely (see text for details).

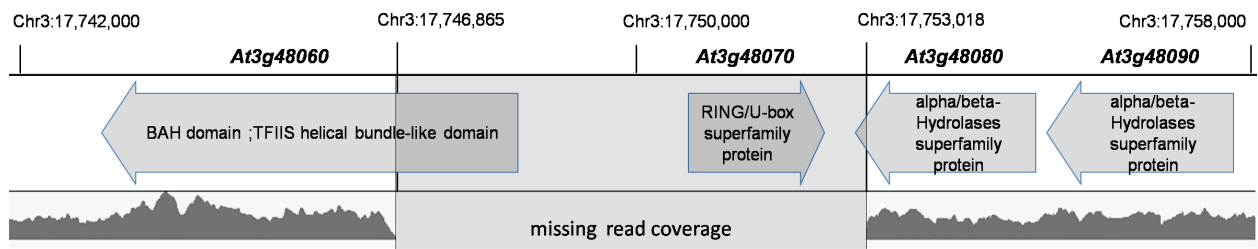


Figure 5: Coverage plot for the large deletion in line GK_144F03

Display of a region on chromosome 3 that is affected by the deletion caused by the T-DNA insertion. The cumulative coverage for both strands is shown and indicates a deletion affecting three genes of which *At3g48070* is lost completely. The mapping data were visualised using Genomeview {Abeel, 2012 #4327}.

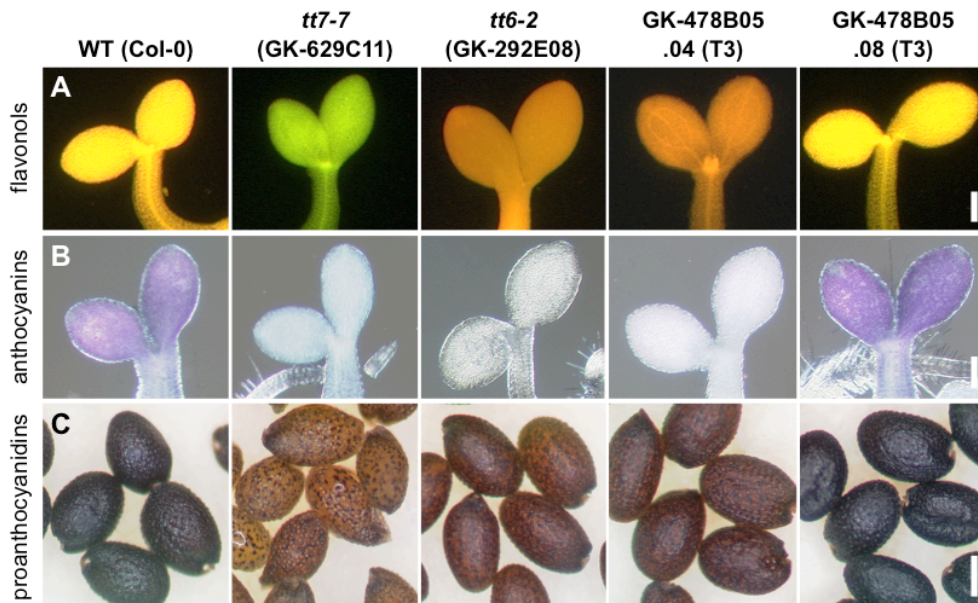


Figure 6: Large deletion in line GK_478B05 affecting *At3g51240* and causing the corresponding *tt6* phenotype

GK_478B05.04 which is homozygous for the insertion shows the same phenotype as the *tt6* knockout GK_292E08. GK_478B05.08 (heterozygous sibling) shows the WT-phenotype. The phenotypical characterization confirms the removal of the gene *At3g51240*. Flavonol accumulation in seedlings was visualised with Diphenyl boric acid 2-aminoethylester (DPBA) and detected using epifluorescence microscopy in homozygous (.04 T3) and heterozygous (.08 T3) plants of GK_478B05 and wild-type as well as *tt6-2* and *tt7-7* plants as controls. Anthocyanins in seedlings were detected using microscopy after bleaching with norflurazon. The proanthocyanidin content of mature seeds was visualised with DMACA staining and detected with a binocular.

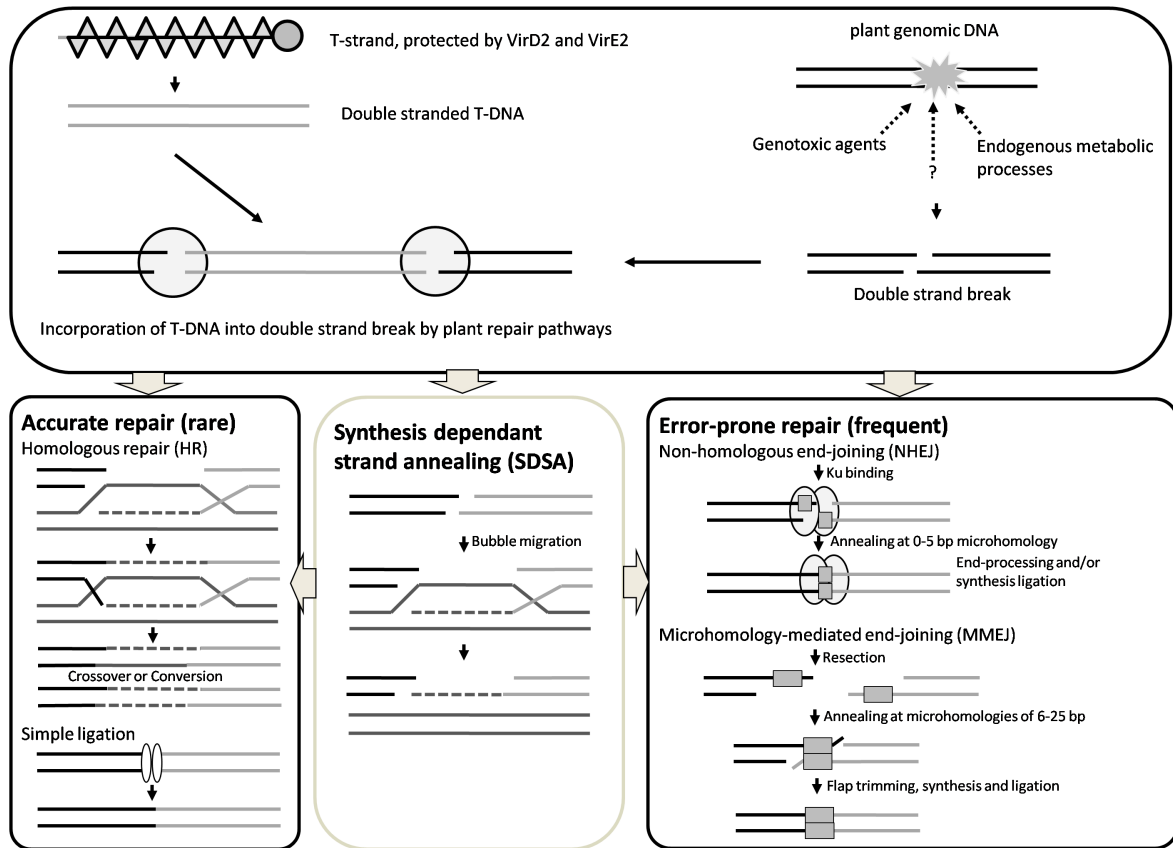


Figure 7: Repair mechanisms used for integration

Summary of double-strand breaks repair mechanisms hypothesized to contribute to T-DNA integration. DSBs occurring for various reasons within the host genome are targets for integration of a double-stranded T-DNA molecule. Integration occurs mainly via the NHEJ and MMEJ pathways, both requiring some microhomology. Accurate repair via simple ligation or HR is very rare and their characteristics almost absent from our data. Before repair takes place, SDSA may create fillers of additional DNA sequence before one of the repair pathways leads to an intact double stranded DNA. This may also explain the presence of inversions at the integration site, when two double-strand breaks occur within close genomic proximity and the fragment gets reincorporated in opposite orientation.

Danksagung

Mein besonderer Dank gilt Prof. Dr. Bernd Weisshaar für die Vergabe des interessanten Promotionsthemas und der Möglichkeit, diese Arbeit am Lehrstuhl für Genomforschung durchführen zu können. Ich bin dankbar für seine Unterstützung und die inspirierenden wissenschaftlichen Diskussionen.

Für die tolle Zusammenarbeit der letzten Jahre möchte ich mich bei dem gesamten GABI-Kat Team und insbesondere bei Dr. Gunnar Huep bedanken, der mich mit Ideen und konstruktivem Feedback immer unterstützt hat. Ebenso danke ich meinen Kollegen vom Lehrstuhl für Genomforschung und aus der Arbeitsgruppe Computational Genomics/BRF für die schöne Arbeits-Atmosphäre, sowie den Studenten, die in Form von Projektmodulen und Bachelorarbeiten an GABI-Kat mitgewirkt haben.

Für die bedingungslose Unterstützung und vielen Aufmunterungen möchte ich besonders meiner Freundin und meinen Eltern danken.

Erklärung

Ich, Nils Kleinbölting, erkläre hiermit, dass ich die Dissertation selbständig erarbeitet und keine anderen als die in der Dissertation angegebenen Hilfsmittel benutzt habe. Alle aus der Literatur entnommenen Zitate sind als solche kenntlich gemacht.

Weiterhin erkläre ich, dass die vorliegende Dissertation weder vollständig noch in Auszügen einer anderen Fakultät mit dem Ziel vorgelegt worden ist, einen akademischen Titel zu erwerben. Ich bewerbe mich hiermit erstmalig um den Doktorgrad der Naturwissenschaften der Universität Bielefeld.

Bielefeld, den 13. April 2015

Nils Kleinbölting