Berichtsblatt

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

3. Titel

GABI-Future - Verbundvorhaben: Eine Sammlung von Doppelmutanten duplizierter Gene bei *Arabidopsis thaliana* (GABI-DUPLO), Teilvorhaben C: HMGU München

4. Autor(en) [Name(n), Vorname(n)]	5. Abschlussdatum des Vorhabens
Dr. Mayer, Klaus	31.12.2010
	6. Veröffentlichungsdatum 9.8.2011
	7. Form der Publikation Bericht
8. Durchführende Institution(en) (Name, Adresse) Helmholtz Zentrum München Inglstädter Landstr. 1 85764 Neuherberg	9. Ber. Nr. Durchführende Institution
	10. Förderkennzeichen 0315055C
	11. Seitenzahl 7
12. Fördernde Institution (Name, Adresse) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. Literaturangaben 1
	14. Tabellen 1
	15. Abbildungen 4

16. Zusätzliche Angaben

17. Vorgelegt bei (Titel, Ort, Datum)
Gabi Statusseminar

18. Kurzfassung

Kurzbeschreibung der Projektfragestellungen

GABI DUPLO beschäftigte sich mit der Erstellung einer Kollektion von Doppelmutanten zur Studie von resultierenden Phänotypen die nicht durch endogene duplizierte Regionen balanciert werden. Um ein rationales experimentelles Design durchzuführen sollte durch in silico Methodiken und Berücksichtigung verschiedener genomischer und funktionaler Parameter eine Vorauswahl geeigneter und vielversprechnder Kandidaten Genpaarungen selektioniert weren und diese dann nachfolgend funktionell und bezüglich evolutionärerer Parameter untersucht und näher charakterisiert werden.

Ziele und Ergebnisse

Der erste Teil der Analysen umfasste die Identifizierung von Kandidaten für GABI-DUPLO, die eine mögliche funktionale Redundanz im Genom von *Arabidopsis thaliana* besitzen. Für diese Studie wurde funktionale Redundanz auf Basis von Sequenz- und Expressionsähnlichkeit ermittelt.
Der zweite Teil der Studie zielte auf die Ermittlung molekularer Grundlagen für genetische Redundanz in duplizierten Genen in *Arabidopsis*. Hierfür wurden konservierte und deletierte cis-regulatorische Elemente in duplizierten Genen betrachtet.
Neben der eng verwandten Art *Arabidopsis lyrata* mit einer Divergenzzeit von etwa 8 Millionen Jahren wurden auch in weitere dikotylen Genomen - Pappel und Wein - Orthologe berechnet. Die Detektion wurden mittels Markow Clustering (orthoMCL) durchgeführt und komplementierten die Syntenie Analysen zwischen den beiden *Arabidopsis* Arten.

19. Schlagwörter
Arabidopsis; Redundanz; duplizierte gene; Doppelmutanten Kollektion

20. Verlag	21. Preis

Abschlussbericht

Zuwendungsempfänger:	Förderkennzeichen:
Helmholtz Zentrum München	**0315055C**
Vorhabenbezeichnung: **GABI-Future - Verbundvorhaben: Eine Sammlung von Doppelmutanten duplizierter Gene bei *Arabidopsis thaliana* (GABI-DUPLO), Teilvorhaben C: HMGU München**	
Laufzeit des Vorhabens: **1.1.2008 – 31.12.2010**	
Berichtszeitraum: **Abschlussbericht**	

Kurzbeschreibung der Projektfragestellungen

GABI DUPLO beschäftigte sich mit der Erstellung einer Kollektion von Doppelmutanten zur Studie von resultierenden Phänotypen die nicht durch endogenen duplizierte Regionen balanciert werden. Um ein rationales experimentelles Design durchzuführen sollte durch in silico Methodiken und Berücksichtigung verschiedener genomischer und funktionaler Parameter eine Vorauswahl geeigneter und vielversprechnder Kandidaten Genpaarungen selektioniert weren unddiese dann nachfolgend funktionell und bezüglich evolutionärerer Parameter untersucht und näher charakterisiert werden.

Ziele und Ergebnisse

Der erste Teil der Analysen umfasste die Identifizierung von Kandidaten für GABI-DUPLO, die eine mögliche funktionale Redundanz im Genom von *Arabidopsis thaliana* besitzen. Für diese Studie wurde funktionale Redundanz auf Basis von Sequenz- und Expressionsähnlichkeit ermittelt. Abb.1 zeigt eine Übersicht über das Schema, mithilfe dessen Kandidaten für GABI-DUPLO selektiert wurden.

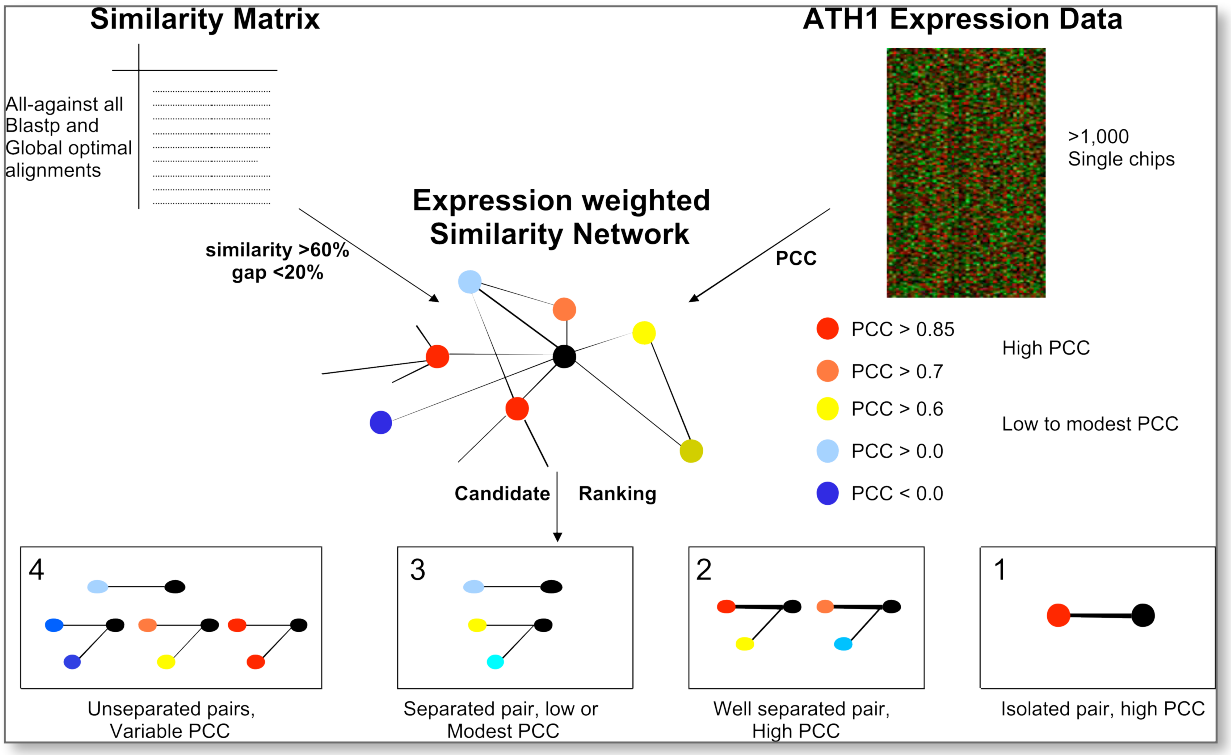


Abb. 1: Selektionsschema für funktional redundante Genduplikationen in Arabidopsis thaliana.

Aus Protein- und Expressionsähnlichkeiten wurden funktional redundante, eng verwandte Genfamilien berechnet. Die Klasseneinteilung in vier Klassen anhand der Expressionsmuster und der Expansion der Genfamilie ist im unteren Bereich der Abbildung illustriert. Weitere Erläuterungen im Text.

Die Identifizierung paraloger Gene in *Arabidopsis thaliana* erfolgte mittels der Berechnung einer Ähnlichkeitsmatrize mit nachgeschalteten Filtern. Im ersten Schritt wurde eine Distanzmatrize für alle Gene anhand ihrer Proteinähnlichkeit mittels des blastp Programms berechnet. Für die folgenden Analysen wurden nur Genpaare betrachtet, die einen minimalen E-value von kleiner gleich 10^{-10} aufwiesen. Optimale globale Alignments wurden für diese Genpaare berechnet und solche Genpaare als eng verwandte Paraloge klassifiziert, die maximal 20% Alignment-Lücken und mindestens 60% Proteinähnlichkeit aufwiesen. Die einzelnen Ergebnisse der paarweisen Analysen wurde in einen ungerichteten Graphen integriert, in dem Gene Knoten und gefilterte Ähnlichkeiten Kanten zwischen zwei Knoten darstellten. Eng verwandte paraloge Cluster bzw. Genfamilien konnten anschließend einfach als „connected components“ ermittelt werden. Diese Cluster enthalten somit sequenz-redundante Genpaare. In einem zweiten Schritt wurden die paarweisen Expressionsdistanzen mittels des Pearson Korrelationskoeffizienten für all Arabidopsis Gene bestimmt. Die Datengrundlage waren hierbei 1765 Arrays inklusive Replikaten, die eine Vielzahl verschiedener Gewebe, Stadien, Reizen und Umweltbedingungen untersuchten. Die globale Korrelation über eine Vielzahl experimenteller Bedingungen selektiert auf Genpaare, die ein gleiches Antwortverhalten für eine breite Auswahl biologischer Prozesse zeigen und deshalb bezüglich ihrer Expression hochredundant sind. Anhand der Expressionsmatrize und der Sequenzähnlichkeiten wurden Genpaare in vier verschiedene Klassen eingeteilt: Klasse 1 und 2 enthielten Gene mit hoher (Pearson Korrelation > 0.7),

Klasse 3 und 4 solche mit geringer Expressionsähnlichkeiten. Klasse 1 und 3 umfassten einzelne Genpaare, Klasse 2 und 4 enthielten expandierte Genfamilien, in denen mehr als zwei Gene mit einer hohen Sequenzähnlichkeit enthalten waren. Um mögliche zusätzliche Redundanzen in diesen expandierten Clustern zu vermeiden, wurden nur Genpaare aus den Klassen 1 und 3 ausgewählt. Eine zusammenfassende Übersicht zeigt Tabelle 1.

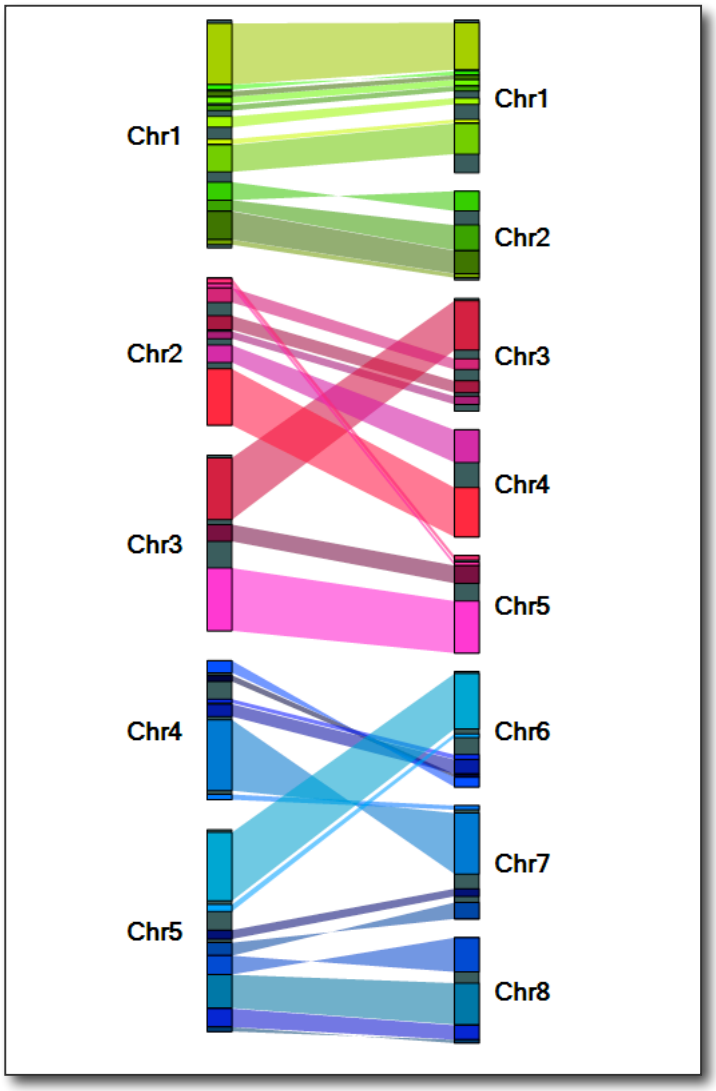
Klasse	Cluster Zahl	GABI-DUPLO
1	1035	487
2	293	0
3	807	378
4	8203	0
Summe	**10338**	**865**

Tabelle 1: Frequenz der paralogen Cluster/Genfamilien pro Kandidatenklasse.

Die erste Spalte zeigt die Klasse, die zweite Spalte gibt die Zahl der paralogen Cluster bzw. Genfamilien für jede einzelne Klasse an, die dritte die Zahl der Cluster, die für GABI-DUPLO ausgewählt wurden. Die Auswahl beschränkte sich auf solche Cluster, für deren Mitglieder Insertionslinien in beiden Genen eines Paares vorhanden waren. Eine Beschreibung der vier Klassen ist im Text angegeben.

In enger Kooperation mit der Gruppe von Bernd Weisshaar wurden mittels der GABI-KAT und SALK Insertionslinien geeignete Kandiatenpaarungen für diese Studie ermittelt. In einem letzten Schritt wurden anschliessend Genpaare, die einen genomischen Abstand von weniger als 2 Mb im Arabidopsis Genom hatten, aus den Kandidaten entfernt. Aufgrund der engen genetischen Kopplung dieser tandem duplizierten Gene ist die Generierung von Doppelmutanten nicht praktikabel. Zusätzlich sollte die Insertion für beide Gene innerhalb des Start- und Stopcodons der kodierenden Regionen liegen, um ein starkes bzw. amorphes Allel für Kreuzungen zur Verfügung zu haben. Das finale Set der Kandidaten für die Doppelmutanten-Analyse umfasste 849 Genpaare.

Der zweite Teil der Studie zielte auf die Ermittlung molekularer Grundlagen für genetische Redundanz in duplizierten Genen in *Arabidopsis*. Hierfür wurden konservierte und deletierte cis-regulatorische Elemente in duplizierten Genen aus der oben angegebenen Analyse betrachtet. Durch die Einbeziehung orthologer Promoteren und Gene kann die Performanz für



die Detektion cis-regulatorischer Elemente wesentlich gesteigert werden. In einem ersten Schritt wurden deshalb in dieser Studie einerseits Syntenien und orthologe Gene zu nah verwandten Genomen von *Arabidopsis thaliana*, namentlich *Arabidopsis lyrata*, berechnet. Abb. 2 illustriert die engen syntenischen Beziehungen zwischen den beiden *Arabidopsis* Arten.

Abb.2 : Syntenie zwischen den 5 Chromosomen von Arabidopsis thaliana (linke Spalte), und den 8 Chromosomen von Arabidopsis lyrata (rechte Spalte) [Hu et al., (2011): Nature Genetics, 476-481].

Neben der eng verwandten Art *Arabidopsis lyrata* mit einer Divergenzzeit von etwa 8 Millionen Jahren wurden auch in weitere dikotylen Genomen - Pappel und Wein - Orthologe berechnet. Die Detektion wurden mittels Markow Clustering (orthoMCL) durchgeführt (siehe Abb. 3) und komplementierten die Syntenie Analysen zwischen den beiden *Arabidopsis* Arten.

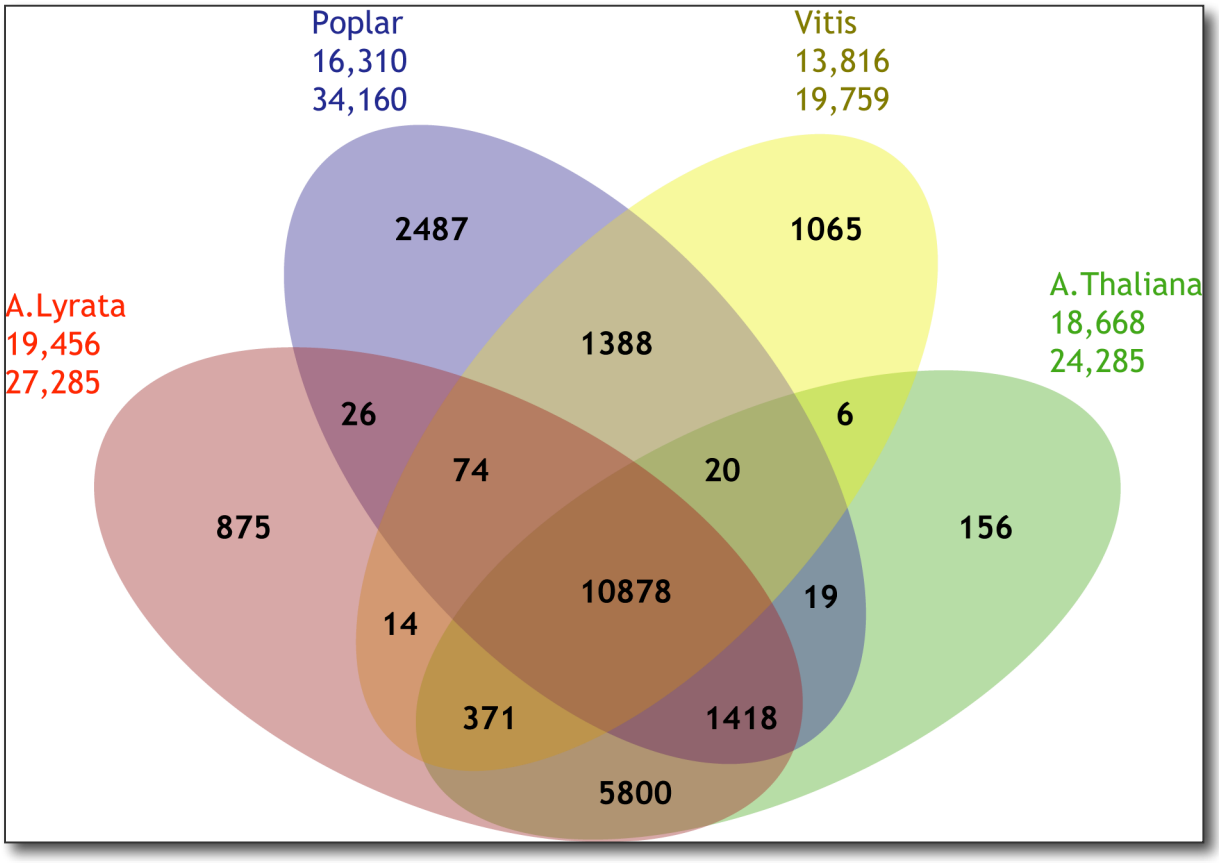


Abb.3 : Orthologe Cluster zwischen vier dikotylen Arten.

Die Abbildung zeigt die orthoMCL Analyse zwischen vier vollständigen Proteomen der Spezies Arabidopsis thaliana, Arabidopsis lyrata, Vitis vinifera (Wein) und Pappel (Populus trichocarpa). Die Zahl der von den Spezies geteilten Cluster bzw. eng verwandten Genfamilien ist in den jeweiligen Bereichen des Venn Diagramms eingetragen, die Zahl in der ersten Zeile unterhalb jeder Spezies gibt die Zahl der Cluster für die Spezies insgesamt an, in der zweiten Zeile darunter ist die Zahl der für das jeweilige Proteom geclusterten Gene gezeigt.

Mithilfe der gefundenen Orthologie-Beziehungen wurden mittels einer „Network level conservation"-Analyse Kandidaten für cis-regulatorische Elemente in den Upstream-Regionen der jeweiligen Orthologen detektiert. Aus den paarweisen Vergleichen wurde für die vier Genome für jeden orthologen Cluster Konsenus Promoterarchitekturen erstellt. Diese werden zur Zeit mit den Daten aus den GABI-DUPLO Set abgeglichen und darauf übertragen, um die möglichen, bisher gefundenen Redundanzen zu erklären. Eine Analyse von Tandem-duplizierten Genen ist zur Zeit in Arbeit. Die Dynamik der Tandem Duplikationen wird mittels gen-basierter Alignments der chromosomalen syntenischen Regionen zwischen *Arabidopsis lyrata* und *thaliana* untersucht. Ein Beispiel für eine solche Region ist in Abb. 4 dargestellt.

RegionID	Evalue	tID	AthGeneID	yID	AlyGeneID
381.387_LHS	300.00	657	AT3G26150	355	fgenesh1_pm.C_scaffold_5000301
381.387_LHS	200.00	657	AT3G26160	355	Al_scaffold_0005_443
381.387_LHS	------		---------	355	Al_scaffold_0005_444
381.387_LHS	200.00	657	AT3G26170	355	fgenesh2_kg.5__376__AT3G26170.1
381.387_LHS	------	657	AT3G26180		---------
381.387_LHS	200.00	657	AT3G26190	355	fgenesh2_kg.5__379__AT3G26190.1
381.387_LHS	200.00	657	AT3G26200	355	fgenesh2_kg.5__380__AT3G26200.1
381.387_LHS	200.00	657	AT3G26210	355	scaffold_500511.1
381.387_LHS	200.00	657	AT3G26220	355	fgenesh2_kg.5__382__AT3G26220.1
381.387_LHS	200.00	657	AT3G26230	355	scaffold_500515.1
381.387_LHS	------		---------	355	fgenesh1_pg.C_scaffold_5000378
381.387_LHS	------		AT3G26235		---------
381.387_LHS	------	871	AT3G26240		---------
381.387_LHS	------	871	AT3G26250		---------
381.387_LHS	109.00	657	AT3G26270	355	scaffold_500517.1
381.387_LHS	------		---------		scaffold_500518.1
381.387_LHS	117.70	657	AT3G26280	355	fgenesh1_pg.C_scaffold_5000380
381.387_LHS	200.00	657	AT3G26290	355	fgenesh1_pg.C_scaffold_5000381
381.387_LHS	200.00	657	AT3G26300	355	fgenesh2_kg.5__386__AT3G26300.1
381.387_LHS	200.00	657	AT3G26310	355	Al_scaffold_0005_456
381.387_LHS	------		---------		Al_scaffold_0005_457
381.387_LHS	200.00	657	AT3G26320	355	fgenesh2_kg.5__388__AT3G26330.1
381.387_LHS	161.70	657	AT3G26330	355	fgenesh1_pg.C_scaffold_5000386

Abb. 4: Gemeinsamer Tandem cluster in *Arabidopsis lyrata und Arabidopsis thaliana*.

Die Abbildung zeigt einen kleinen Ausschnitt der syntenischen Region 381.387_LHS zwischen Arabidopsis thaliana und A. lyrata. Die Synteny beruht auf einem optimalen globalen Alignment der einzelnen vorhergesagten Gene beider Genome. Die Gene eines Tandem Clusters in Arabidopsis thaliana mit der Cluster ID (tID) 657 sind orange, Gene eines Tandem Clusters in A. lyrata mit der Cluster ID (yID) 355 sind grün unterlegt. Deutlich ist die Orthologie beider Cluster erkennbar, die ebenfalls duch hoch signifikante Proteinähnlichkeiten (Evalue) belegt sind. Ein weiterer Tandem Cluster in A. thaliana, blau unterlegt, ist in der Genomsequenz eingeschoben und fehlt in A. lyrata. Analysen wie diese unterstützen die Aufklärung komplexer Orthologie-Beziehungen in höheren Pflanzen, die häufig von einfachen 1:1 Beziehungen abweichen.

Einhaltung der Zeit- und Kostenplanung

Das Projekt umfasste die unten gelisteten einzelnen Arbeitsschritte und Analysen

- Berechnung und Klassifikation duplizierter Gene anhand aktueller Annotation des *Arabidopsis thaliana* Genoms; Korrelationsanalysen der Expression duplizierter Gene und Kandidatenauswahl.
- Detektion orthologer Beziehungen anhand von Ähnlichkeitsmatrizen und Syntenie zwischen den Genomen von *Arabidopsis thaliana, Arabidopsis lyrata* und *Capsella rubella*.
- Vergleichende Promoterstudien und Detektion signifikanter Motive anhand von Coexpression und phylogenetischer Verwandtschaft.
- Erstellen transkriptioneller Netzwerke und Module.
- Synthese der bioinformatischen Ergebnisse mit experimentellen Ergebnissen der Gruppen von P1 und P2.

Im Schritt zwei konnten keine Syntenie und Ähnlichkeitsanalysen zum Genom von Capsella rubella durchgeführt werden. Im Gegensatz zur ursprünglichen Projektvoranahme wurde das

Genom bisher noch nicht fertiggestellt. Die Verzögerung war jedoch alusserhab des Einflussbereichs der GABI Duplo Projektpartner (externe Kollaboration). Um dies auszugleichen wurden jedoch Syntenie- und Ähnlichkeitsanalysen gegen zusätzliche Genomreferenzen -Wein und Pappel- durchgeführt. Durch die leicht verzögerte Erstellung von Doppelmutantendaten konnte keine Fusion von bioinformatischen Ergebnissen und experimentellen Daten durchgeführt werden. Nach Verfügbarkeit wird diese Analyse jedoch noch durchgeführt werden.

Kostenplanung und Verwendung der Projektmittel entsprach der ursprünglichen Planung.

Referenz:

Tina T Hu, Pedro Pattyn, Erica G Bakker, Jun Cao, Jan-Fang Cheng, Richard M Clark, Noah Fahlgren, Jeffrey A Fawcett, Jane Grimwood, Heidrun Gundlach, Georg Haberer, Jesse D Hollister, Stephan Ossowski, Robert P Ottilar, Asaf A Salamov, Korbinian Schneeberger, Manuel Spannagl, Xi Wang, Liang Yang, Mikhail E Nasrallah, Joy Bergelson, James C Carrington, Brandon S Gaut, Jeremy Schmutz, Klaus F X Mayer et al. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature Genetics*, **43**, 476–481.