Zweite molekulargenetische Charakterisierung von Süßkirschengenotypen der Deutschen Genbank Obst mittels eines hochauflösenden SNP Marker Arrays

Abschlussbericht

Auftragnehmer	SGS Institut Fresenius GmbH TraitGenetics Section (SGS IF TG) Am Schwabeplan 1b 06466 Seeland OT Gatersleben Germany
Projektnummer	**2821BE001**
Laufzeit	20.07.2021 – 10.12.2021
Berichtszeitraum	20.07.2021 – 10.12.2021
Autoren	Dr. Jörg Plieske, Dr. Anja Hohmeyer, Dr. Heike Gnad, Dr. Martin Ganal

zuletzt korrigiert durch den Projektträger BLE am 11.01.2022

1. Ziele und Aufgabenstellung des Projekts

Ziel des Projekts ist die molekulargenetische Charakterisierung von weiteren 184 Süßkirschengenotypen der Deutschen Genbank Obst (DGO) mittels des hochauflösenden Infinium RosBREED Cherry_V2 Genotyping Arrays, einschließlich der dafür notwendigen Isolierung der genomischen DNA aller Genotypen. Die Daten sollen anschließend zusammen mit den Genotypisierungsdaten des vorangegangenen Projektes (FKZ: 2819BE001) zur Identifizierung möglicher Duplikate genutzt werden.

2. Planung und Ablauf des Projekts

Die für die Genotypisierung notwendigen Blattproben wurden SGS IF TG zu Beginn des Projektes von der Koordinierungsstelle der DGO zur Verfügung gestellt. Gleichzeitig erhielt SGS IF TG eine Liste (Excel-Format) mit Informationen zu den Proben, in der folgende Daten enthalten waren: Sortenname, Nummer der molekularen Gruppe, Akzessionsnummer, Standort des Baumes und Probennummer. Zunächst erfolgte die Isolierung der genomischen DNA aller Genotypen. Die Genotypisierung der Kirschensorten (mit Kontrollproben der SGS IF TG) der DGO erfolgte mit dem Infinium RosBREED Cherry_V2 Genotyping Array. Die von SGS IF TG verwendeten Kontrollen dienten der Überprüfung der Robustheit der Marker und der Reproduzierbarkeit der Ergebnisse. Das Cluster File für die Auswertung der Daten wurde in einem vorangegangenen Projekt (FKZ: 2819BE001) entwickelt und nicht modifiziert. Im Anschluss wurden die Rohdaten zusammen mit den Rohdaten des vorangegangenen Projektes mit Hilfe des etablierten Cluster Files ausgewertet und eine Genotyptabelle im IUB Code erstellt.

Die Genotyptabelle enthält den Sortennamen, die Nummer der molekularen Gruppe, die Akzessionsnummer, die Standortnummer des Baumes sowie die Probennummer und die Analysedaten für die ausgewählten Marker. Alle Proben wurden nach identischen Fingerprints sortiert und jedem einzelnen Fingerprint eine eindeutige Identifikationsnummer ID zugeordnet. Diese eindeutige ID wurde in einer gesonderten Zeile ebenfalls in der Tabelle aufgeführt. Mögliche Duplikate wurden in dieser Tabelle sichtbar gekennzeichnet (farblich hinterlegt). Die Übergabe der Ergebnisse erfolgt auf elektronischem Wege (als MS-Excel Tabelle).

3. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Array-basierte Genotypisierung mit molekularen Markern, wie beispielsweise SNPs (single nucleotide polymorphisms), bietet die Möglichkeit viele tausend Marker im Hochdurchsatz, reproduzierbar und kosteneffizient zu untersuchen. Die hier genutzte Illumina Infinium Technologie basiert auf bewährter Chemie kombiniert mit der zuverlässigen Bead Array-Plattform und sorgt so für eine hohe Datenqualität, hohe Call-Raten sowie eine hohe Reproduzierbarkeit. Der Infinium RosBREED Cherry_V2 Genotyping Array bzw. sein Vorgängerarray RosBREED cherry 6K SNP array v1 wurde für die diploide Süßkirsche (*Prunus avium*) und allotetraploide Sauerkirsche (*P. cerasus*) entwickelt. Die ausgewählten SNPs basieren auf der Next-Generation Sequenzierung von diversem genetischem Material, die 25 Milliarden Basenpaare (Gb) Kirschensequenzen zur Verfügung stellte. Basierend darauf wurden genomweite SNPs für die Süßkirsche und für die beiden Subgenome der Sauerkirsche (Avium Subgenom und Fruticosa Subgenom) identifiziert. Da der genutzte Array bereits vom Auftraggeber ausgewählt wurde, wird hier nicht weiter im Detail auf die Zusammensetzung und Qualität des Arrays eingegangen.

SGS IF TG wurde im Jahr 2000 gegründet und arbeitet somit seit über 20 Jahren erfolgreich auf dem Gebiet der Pflanzengenotypisierung. SGS IF TG entwickelt nicht nur SNPs, sondern bietet die routinemäßige Hochdurchsatzgenotypisierung als Dienstleistung an. Unsere Erfahrung und Expertise in der Auswertung derartiger Daten wird regelmäßig durch zahlreiche wissenschaftliche Publikationen demonstriert.

4. Material und Methoden

a) DNS Extraktion

Die DNS Extraktion wird im 96 Well-Format durchgeführt und beruht auf einer CTAB-Methode.
CTAB-Puffer (500ml):

- 2% CTAB (Cetyltrimethylammoniumbromid) (10g)
- 200mM Tris (100ml 1M Tris pH8)
- 20mM EDTA (20ml 0,5M Na-EDTA pH8)
- 1,4M NaCl (40,9g)
- 1% PVP (5g 40.000g/mol)

Zu 500 ml CTAB-Puffer werden frisch 1,9 g Na-Bisulfit hinzugegeben und die Lösung auf 60°C vorgewärmt. Das Blattmaterial wird in 96er deep-well Mikrotiterplatten mit Kugeln unter Stickstoff vermahlen, dann 400 µl warmer Puffer zugegeben, homogenisiert und 15 min bei 60°C im Wasserbad inkubiert. Im Anschluss findet eine Chloroform/Isoamyl (24:1) Aufreinigung statt. Dafür werden 400 µl Chloroform/Isoamyl (24:1) zugegeben, homogenisiert und 20 min bei 3.000g zentrifugiert. 115 µl Überstand werden in eine neue Platte überführt und mit 85 µl Isopropanol gefällt. Nach der Zentrifugation (45 min, 3.500g) wird der Überstand abgegossen, das Pellet getrocknet und in Wasser gelöst.

b) Genotypisierung mit Infinium RosBREED Cherry_V2 Genotyping Array

Die Details der Genotypisierung können dem im Anhang befindlichen Infinium® HD Assay Ultra Protocol (Guideinfinium_hd_ultra_user_guide_11328087_revb Illumina HD Protokolls) von Illumina entnommen werden.

Der Array basiert auf dem Infinium HD-Typ und kann parallel 13.559 Marker analysieren. Diese setzen sich aus Markern speziell für die Süßkirsche und aus sauerkirschspezifischen Markern zusammen.

c) Clusterfileentwicklung und Auswertung der Daten

Die Rohdaten aus dem gescannten Array werden mit Hilfe der Genomstudio Software ausgewertet (User Guide befindet sich im Anhang). Das Clusterfile dient zur Qualitätskontrolle der Marker und als Schablone für eine reproduzierbare Bestimmung der Allele in einzelnen Experimenten. Auf dieser Basis werden anschließend die Genotypendaten generiert, die mit Bezug zur Quellsequenz im IUB Code ausgegeben werden.

d) Duplikat-Analyse

Die Duplikat-Analyse erfolgte mit einem eigens bei SGS IF TG für die Haplotypen- und Fingerprintanalyse entwickelten Skript. In diesem Skript werden die genetischen Fingerprints der Sorten über alle Marker verglichen und jedem unterschiedlichen Fingerprint eine eindeutige ID zugewiesen. Genotypen (Sorten) mit gleichem Fingerprint tragen die gleiche ID. Zusätzlich wurde eine Verwandtschaftsanalyse mit NTSYS analog zum vorangegangenen Projekt durchgeführt.

5. Ausführliche Darstellung der Ergebnisse

Die Probenübergabe an SGS IF TG erfolgte am 13.10.2021 zusammen mit allen nötigen digitalen Informationen. Im Anschluss erfolgte die DNS Extraktion. Eine detaillierte Qualitätskontrolle der DNS-Extraktion ist dem Anhang zu entnehmen (PCS21001 Layout).

Die Genotypisierung erfolgte in zwei 96-Well Mikrotiterplatten nach dem im Anhang befindlichen Protokoll des Illumina Infinium HD Workflows „Infinium® HD Assay Ultra Protocol“. Die Rohdaten wurden mit Hilfe der Software Genome Studio 2.0 ausgewertet. Das verwendete Cluster File war in dem vorangegangenen Projekt zur Analyse des ersten Süßkirschen Sets entwickelt worden (Az.: 123-

02.05-20.0129-19-II-G) und wurde in diesem Projekt nicht modifiziert. Insgesamt wurden 6.356 polymorphe Marker identifiziert. Die Anzahl der auch später in der weiteren Analyse der Daten verwendeten polymorphen Marker war erheblich höher als in der ersten Analyse (3855). In der vorliegenden Auswertung beider Datensätze wurden auch Marker mit geringer Minor Allele Frequency (MAF) hinzugezogen, um eine möglichst feine Unterscheidung der Genotypen zu ermöglichen. Die MAF der Marker lag zwischen 0.500 und 0.0027 (Abb. 01).

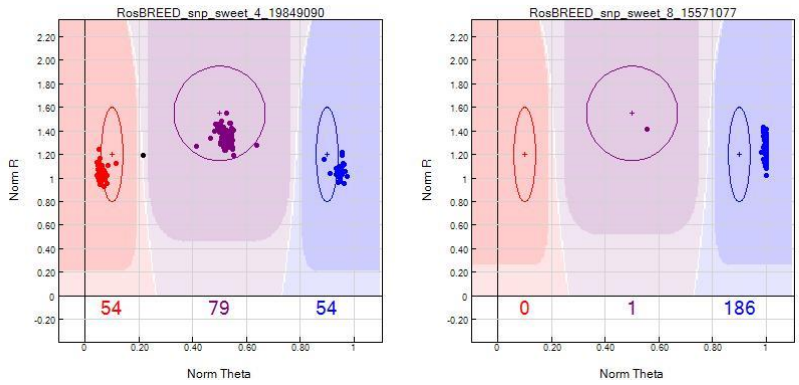

Abb. 01: Beispiele der Clusteranalyse der 192 Samples inklusive biologischer und technischer Kontrollen. Links: MAF 0.5, rechts: MAF 0.0027

Die Genotypdaten wurden routinemäßig mit Hilfe der von SGS IF TG entwickelten Workbench ausgelesen und im Excelformat gespeichert. Für die Identifizierung der Duplikate wurden beide Datensätze, d.h. der Datensatz aus dem vorangegangenen Projekt und der aus dem vorliegenden Projekt, zusammengeführt. Letztendlich umfasste der Datensatz für die Duplikat-Identifizierung 541 Genotypen - wobei für eine Akzession keine Daten generiert werden konnten - und 11 biologische Kontrollen der Akzession „Areko“. Die 541 Genotypen setzten sich wie folgt zusammen: 357 Genotypen aus der ersten Analyse, (inklusive 5 DNS Duplikate für die Qualitätskontrolle der Duplikat-Identifizierung) und 184 Genotypen der zweiten Analyse.

Abb. 02: Ausschnitt aus der Genotypentabelle mit gelber Markierung der Samples in jeweils identischen Haplotypen

Die Identifizierung der Duplikate wurde mit einem von SGS IF TG entwickelten Skript zur Haplotypenidentifizierung durchgeführt. Das Skript wurde für diese Analyse eigens modifiziert und optimiert. Es identifiziert über alle Marker den entsprechenden genetischen Fingerabdruck jeder Akzession. Dieser Fingerabdruck wird anschließend zwischen allen Akzessionen abgeglichen,

identische Muster identifiziert und einem Haplotypen zugeordnet. Jedem unterscheidbaren Haplotyp wird eine eindeutige Bezeichnung – Haplotypen ID – zugewiesen. Die Haplotypen ID jeder Akzession wurde dann in der Genotypentabelle vermerkt und die Daten entsprechend sortiert. Akzessionen mit gleicher Haplotypen ID wurden farbig markiert (Abb. 02). Insgesamt wurden 347 unterschiedliche Haplotypen identifiziert.

In der ersten Analyse wurde die Duplikat-Identifizierung noch ausschließlich mit Hilfe der kommerziellen Software NTSYS durchgeführt. Da diese Software keine befriedigende Schnittstelle für den Export der Daten aufwies, die dann anschließend eine elektronische Verarbeitung der Daten für die Duplikate-Identifizierung ermöglichte, wurde nach einer neuen Lösung dafür gesucht. SGS IF TG hatte im Laufe der letzten Jahre für die Entwicklung optimierter Arrays bereits ein Tool entwickelt, mit der die Redundanz von molekularen Marker-Daten in einem Set von Genotypen elektronisch basiert identifiziert werden konnte. Für die Identifizierung der Duplikate in einem Set von Genotypen mussten im Prinzip nur die Koordinaten umgedreht werden – nicht die Marker waren Gegenstand der Analyse sondern die Genotypen. Anhand der DNS Duplikate und der vielfachen Analyse der Kontrolle „Areko“ konnte die Vertrauenswürdigkeit dieser Vorgehensweise nachgewiesen werden.

In der ersten Analyse wurden in 352 untersuchten Genotypen 303 Haplotypen identifiziert. Das entspricht einem Koeffizienten von 0,86 Haplotypen je Genotyp. In dem gesamten Datensatz von 536 Genotypen wurden 347 Haplotypen identifiziert. Das entspricht einem Koeffizienten von 0,65. Dieser erheblich geringere Koeffizient kann nur dahingehend interpretiert werden, dass das zweite Set Genotypen in signifikanter Verwandtschaft zum ersten Set gestanden hat.

Der technische Fehler bei der Infinum Array Analyse ist erfahrungsgemäß sehr klein. Auch in diesem Projekt wurden alle technischen und biologischen Duplikate zweifelsfrei identifiziert. Dennoch können auch kleine Abweichungen zwischen sehr nahe verwandten Akzessionen dazu führen, dass die Genotypen unterschiedlichen Haplotypen zugeordnet werden. Daher wurden weitere Haplotypenanalysen mit leicht reduzierter Stringenz (0,995 und 0,990) durchgeführt und die Ergebnisse in die Genotypentabelle eingebunden. Auch wurde eine Verwandtschaftsanalyse mit NTSYS analog zum vorangegangen Projekt durchgeführt (Abb. 03). Zusammen ergibt sich mit diesen unterschiedlichen Ansätzen die Möglichkeit, auch nahe verwandte Akzessionen zu identifizieren und ggf. anhand weiterer (pomologischer) Informationen einem Haplotyp zuzuordnen. Da uns diese zusätzlichen Informationen nicht zur Verfügung stehen, ist eine diesbezüglich weitere Analyse des Datensatzes nicht möglich. Für Rückfragen stehen wir jedoch sehr gerne jederzeit zur Verfügung.

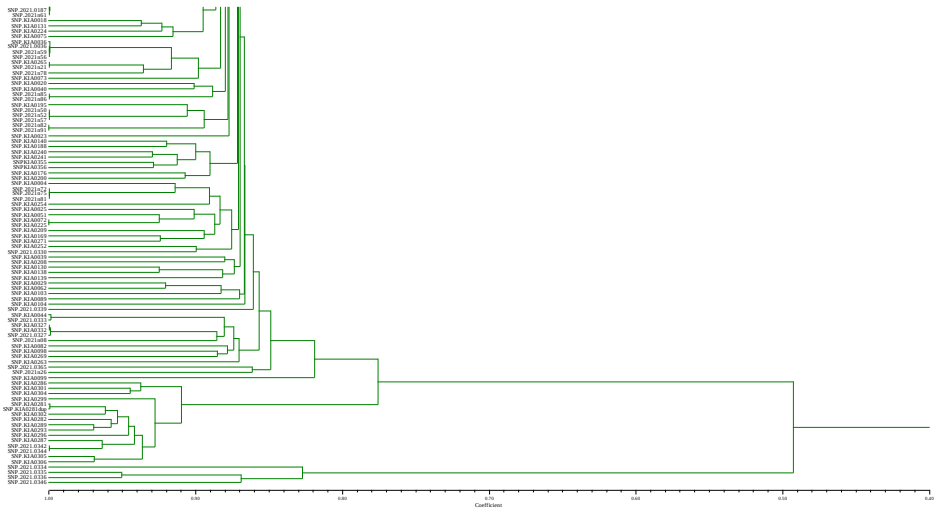

Abb. 03: Ausschnitt aus der Verwandtschaftsanalyse der Akzessionen mit NTSYS (Similarity Coefficient DICE, hierarchical clustering method UPGMA, SAHN algorithm)

6. Voraussichtlicher Nutzen und Verwertbarkeit der Ergebnisse

Insgesamt konnten die Arbeiten erfolgreich durchgeführt werden. Für die Süßkirsche wurden qualitativ sehr hochwertige und reproduzierbare Ergebnisse erzielt und die Duplikate zuverlässig identifiziert. Den voraussichtlichen Nutzen und die genaue Verwertbarkeit der Ergebnisse kann nur vom Auftraggeber beurteilt werden, da die genauen Eigenschaften des untersuchten Materials SGS IF TG unbekannt sind.

7. Zusammenfassung

Ziel des Projekts war die molekulargenetische Charakterisierung von Süßkirschen der Deutschen Genbank Obst (DGO) mittels des Infinium RosBREED Cherry_V2 Genotyping Arrays zur Identifizierung von Duplikaten.

Der Array war für die Charakterisierung der Süßkirschenproben sehr gut nutzbar. Letztendlich wurden 6.356 polymorphe der 13.559 auf dem Array befindlichen Marker für die Analyse der Süßkirschen Akzessionen verwendet. Mit Hilfe dieser Genotypendaten wurde eine Haplotypen- oder genetischen Fingerprintanalyse zur Identifizierung von Duplikaten erfolgreich durchgeführt. Die 536 Süßkirschenakzessionen wurden 347 Haplotypen zugeordnet, wobei eine Akzession nicht ausgewertet werden konnte. Entsprechend wurden 188 Akzessionen als Duplikate identifiziert.

Alle biologischen und technischen Kontrollen wiesen einen eindeutigen Genotyp auf und zeigten, wie stabil und reproduzierbar die arraybasierte Genotypisierung ist.

8. Gegenüberstellung der ursprünglich geplanten zu den tatsächlich erreichten Zielen; ggf. mit Hinweisen auf weiterführende Fragestellungen

Ziel des Projekts war die molekulargenetische Charakterisierung von Süßkirschengenotypen der Deutschen Genbank Obst (DGO) zur Identifizierung möglicher Duplikate. Erreicht wurden eine technisch und qualitativ hochwertige molekulargenetische Charakterisierung der Süßkirschenproben mittels des Illumina Infinium Arrays und die sichere Identifizierung der Duplikate mittels eigens für solche Fragestellungen entwickelter Skripte.

9. Literaturverzeichnis

a) Peace, C; Bassil, N; Main, D; Ficklin, S; Rosyara, U R; Stegmeir, T; Sebolt, A; Gilmore, B; Lawley, C; Mockler, T C; Bryant, D W; Wilhelm, L; Lezzoni A (2012). Development and Evaluation of a Genome-Wide 6K SNP Array for Diploid Sweet Cherry and Tetraploid Sour Cherry. PLOS ONE; December 2012; Volume 7;Issue 12
b) GenomeStudioTM Genotyping Module v1.0 User Guide
c) Infinium® HD Assay Ultra Protocol Guide