Selektion und Evolution

Allelfrequenzschätzungen machen historische Selektionsereignisse sichtbar

JENS LÉON[1], AGIM BALLVORA[1], MICHAEL SCHNEIDER[1,2]
[1] INSTITUT FÜR NUTZPFLANZENWISSENSCHAFT UND RESSOURCENSCHUTZ, UNIVERSITÄT BONN
[2] DEPARTMENT NUTZPFLANZENWISSENSCHAFTEN, FORSCHUNGSINSTITUT FÜR BIOLOGISCHEN LANDBAU (FIBL), FRICK, SCHWEIZ

Allele frequency shifts can result from adaptation or selection and indicate strategies for coping with stress scenarios. Observing these requires genotyping of hundreds of lines individually or in a pooled sample – both rather costly, especially for species with large genomes. Constructing virtual haplotypes from SNP allele frequencies can drastically reduce genotyping time and costs in pooled sampling. Further, we validated three commonly used genotyping strategies for poolseq in crop species.

DOI: 10.1007/s12268-023-1993-9

Die Interpretation der genetischen Diversität in Populationen ist Bestandteil diverser Wissenschaftsfelder und erlangt zunehmend in den angewandten Wissenschaften wie der Landwirtschaft und Züchtung an Bedeutung.

Pflanzenpopulationen sind ortstreu und müssen sich daher an die jeweilige Umgebung anpassen. Moderne Genomanalysen sind in der Lage diejenigen Muster zu erkennen, die die Adaptation an die Umgebung in der genetischen Struktur der Individuen hinterlassen hat. Durch die Selektionsvorgänge gegen oder zugunsten relevanter Eigenschaften haben sich die Allelfrequenzen der betroffenen Genregionen verändert. Je kürzer bestimmte Ereignisse zurückliegen oder je heftiger die Selektionswirkung war, desto ausgeprägter sind die Adaptationsmuster in der Population zu erkennen. Diese Regionen, die *selective sweeps* genannt werden, geben Aufschluss über die Anpassung an veränderte Umweltbedingungen und können somit als Methode zur Identifikation von Kandidatengenregionen verstanden werden (Genotyp-Umwelt-Assoziationen).

Um die *selective sweeps* erkennen zu können, sind präzise Messungen der Allelfrequenzen in den Populationen notwendig. Da die Genotypisierung von einzelnen Individuen zur Allelfrequenzschätzung aufwändig und teuer ist, bietet die Pool-Sequenzierung den Vorteil der Skalierung mit der Anzahl der zu testenden Proben, ohne dass Kosten stark steigen. Obwohl die Pool-Sequenzierung bereits zu einer deutlichen Zeit- und Kostenreduktion gegenüber der Einzelpflanzen-Genotypisierung führt, kann die notwendige hochauflösende Sequenzierung Projektbudgets trotzdem übersteigen, da unsere Kulturarten regelmäßig große Genome besitzen. Kostengünstige Sequenziermethoden führen zumeist zu einer geringeren Abdeckung des Genoms. Erschwerend kommt hinzu, dass der Genomaufbau insbesondere bei Kulturarten – durch beispielsweise Autopolyploidie oder repetitive Sequenzen – Probleme bereiten kann.

Haplotypfrequenz statt Allelfrequenz

Für Pool-Sequenzierungen werden üblicherweise Abdeckungsraten von 50–100x und mehr empfohlen. Um die Kosten weiter zu senken, könnte diese Sequenziertiefe redu-

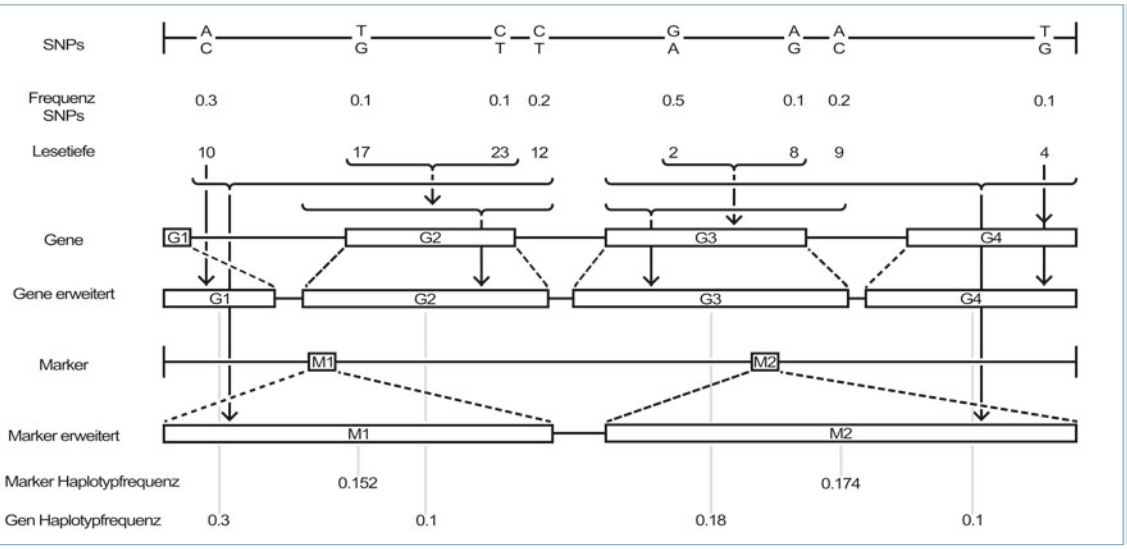


Abb. 1: Von der SNP zur Haplotyp-Allelfrequenz. Grafische Veranschaulichung der Gen-, Marker- oder Contig-annotierten Haplotypen.

ziert werden. Bei einer Abdeckung von unter 50 Sequenzen pro Base ist allerdings eine Angabe der Allelfrequenz eines *single nucleotid polymorphismus* (SNP) unzuverlässig [1, 2].

Bei der Auswertung kann man sich allerdings zu Nutze machen, dass die zahlreichen SNP auf dem Genom strukturiert vorliegen. Viele von ihnen sind miteinander gekoppelt und nur voneinander getrennt, wenn Kopplungsbrüche bzw. *crossing over*-Ereignisse während der jüngsten Meiosen zwischen diesem Paar von SNPs stattgefunden haben. Somit tragen die SNPs aus genetischer Sicht de facto dieselbe Information. Addiert man nun die Sequenzen dieser beiden SNPs zusammen, lässt sich die Lesetiefe des Haplotypen, der aus den beiden SNP gebildet wird, verdoppeln. Summiert man die Lesetiefe von zehn SNP bei einer durchschnittlichen Abdeckung von zehn Sequenzen pro SNP zusammen, ergibt sich eine Haplotyp-Sequenztiefe von 100x – ausreichend zur genauen Bestimmung der Allelfrequenz. Möglich macht es die Tatsache, dass Kopplungsbrüche über das Genom gesehen verhältnismäßig selten sind. Die Zahl der Kopplungsbrüche ist allerdings nicht statisch. Sie ist u. a. abhängig von der spezifischen Region in den jeweiligen Chromosomen, dem Genomaufbau und den Umweltbedingungen [3]. Es bleibt gleichwohl festzuhalten, dass Kopplungsbrüche selten sind. Gehen wir in der hier getesteten Gerste beispielhaft von 3–5 Rekombinationen (also erkennbare Kopplungsbrüche) pro Chromosom und Individuum aus, ist in einer Poolprobe von 300 Genotypen entsprechend mit etwa 900–1.500 pro Chromosom zu rechnen. Auf eine durchschnittliche Chromosomengröße bei Gerste von 600 Megabasen (MB) bezogen, hieße das alle 4 bis 7 Kilobasen eine Rekombination.

Die Haplotypen können auf verschiedene Arten konstruiert werden. Neben der Möglichkeit ganze Genomabschnitte in „Haplotyp-Contigs zu clustern, gibt es einen selektiven Ansatz, in dem bekannte Gene oder Marker als Anker für die Haplotypen genutzt werden, womit eine funktionale Analyse angeschlossen werden kann. Nachteilig an diesem Ansatz ist, dass mit der Reduktion auf die Genregionen viele SNPs nicht zugeordnet werden können, da sie außerhalb der Gene liegen (**Abb. 1**). In der Tat ist der Großteil der SNP außerhalb der Gene zu verorten, weshalb es vorteilhaft ist, die Regionen zwischen den Genen ebenfalls zu nutzen. Der von uns gewählte Ansatz erweitert dabei die Start- und Endbase eines Gens in die Inter-Genregion so weit, dass die Lücke zu 90 Prozent geschlossen wird. Somit können die meisten SNPs in der Nähe der Gene auch diesen noch zugeordnet werden, wodurch die Lesetiefe von durchschnittlich 9 je SNP auf 963 je Gen-Haplotyp erhöht werden kann.

Neben der Zusammenführung von SNPs in Haplotypfenstern ist ein nicht minder wichtiger Aspekt die korrekte Zuordnung der Allele der einzelnen SNPs zueinander. Woher soll bekannt sein, ob die drei Basenpaare *G-A*, *A-G* & *A-C* auf den Genomposition 10, 45 und 120 zu den Haplotypen *GAA*, *GGA*, *GGC*, *GAA*, *GAC*, *AAA*, *AGA*, *AGC oder AAC* zusammengefasst werden müssen? Hilfreich wäre hierbei, wenn eine Referenz bekannt ist: Sie könnte Aufschluss darüber geben, welcher der neun möglichen Haplotypen existent ist und welcher höchstwahrscheinlich nicht. Im Fall der Züchtung kann diese Referenz beispielsweise durch ein oder mehrere Eltern gebildet werden, die zur Erstellung der Population miteinander gekreuzt wurden. Sequenziert man diese im Zuge der Analyse mit, ist ihr Haplotyp bekannt – vorausgesetzt, es handelt sich um homozygote Individuen. Bei heterozygoten Eltern ist ein *phasing* notwendig, um die beiden Haplotypen je Elter unterscheiden zu können. Dies ist beispielsweise durch das Sequenzieren mit Long-Reads wie Hi-Fi oder Nanopore möglich.

Validierung des Haplotypisierungsansatzes und Vergleich verschiedener Sequenzieransätze

Die Sequenziermethodik kann ebenfalls einen signifikanten Einfluss auf die Präzision der Allelfrequenzschätzung aus der Pool-Sequenzierung haben. Um diesen zu untersuchen, wurden drei Pools, bestehend aus jeweils knapp 300 Sommergersten Linien einer biparental erzeugten und unselektierten BC_2F_{23} Population jeweils mittels *genotyping by sequencing* (GBS), einer Ansequenzierung der Gentranskripte (RNA) (MACE) und *whole genome sequencing* (WGS) genotypisiert. Hierfür wurden Sequenziertiefen für jede Poolprobe von 30x (GBS, MACE) und 10x (WGS) genutzt. Bei der Zusammenstellung der Pools wurde darauf geachtet,

◀ **Abb. 2:** Gegenüberstellung der drei genutzten Methoden zur Genotypisierung und die Prozessierung der jeweils generierten DNA-Fragmente.

dass alle Individuen eine identische Menge Blattmaterial beisteuern. Zur Validierung wurden alle Individuen der Poolproben mittels 21 *kompetitive allele specific PCR*(KASP)-Markern genotypisiert. Bei der Auswertung wurden die Vergleiche für die Einzel-SNPs, ebenso wie die Gen- und Marker-Haplotypen verglichen (vergl. **Abb. 2**). Zusätzlich wurde der von Tilk *et al.* beschriebene Contig-Haplotyp-Ansatz in den Vergleich mit einbezogen [4].

Ergebnisse

Der Vergleich von Einzelgenotyp-Analyse und Pool-Sequenzierung zeigte, dass mit zunehmender Lesetiefe je Haplotyp die Genauigkeit der Allelfrequenzschätzungen über alle drei Genotypisierungsmethoden zunahm. Die deutlichste Verbesserung wurde dabei beim WGS erzielt, welche gleichzeitig die höchste Anzahl entdeckter SNPs besaß (ca. 4 Mio. SNPs; **Tab. 1A**). Verglichen dazu konnten mit GBS nur etwa 82.000 und mit MACE 13.000 SNPs erfasst werden.

Auffällig war, dass sich bei der GBS-Pool-Sequenzierung bei weniger als 450 Sequenzen pro Haplotyp keine Korrelation zu den KASP-Markern ergab. Die technischen Duplikate, die im Prozess der Sequenzierung entstehen, dürften hierfür ursächlich sein. Während diese Duplikate in der Expressionsanalyse (MACE) und WGS leicht herausgefiltert werden können, ist diese Filterung durch die vorgegebenen Restriktionsschnittstellen nicht zu gewährleisten. So vermischen „echte“ wie technische Duplikate und sorgen für eine verzerrte Frequenzschätzung. Dies sorgt wohl ebenfalls dafür, dass die drei Wiederholungen über alle drei Sequenzier-basierten Genotypisierungsmethoden am deutlichsten voneinander abweichen. Hervorzuheben ist, dass das Entfernen der technischen Duplikate von enormer Bedeutung für die Güte der Pool-Allelfrequenzgenauigkeit ist.

Die MACE-Pool-Seq weist nur eine geringe Anzahl entdeckter Polymorphismen auf, was zum einen an der speziellen Methode der 3'-End-Sequenzierung liegt, zum anderen an der selektiven Genexpression. Ein direkter Einfluss der Genexpression auf die Allelfrequenzschätzung konnte nicht entdeckt werden; da die Replikate sich allerdings auf dem Gen-Haplotyp-Level noch signifikant unterschieden, kann ein solcher Einfluss nicht ausgeschlossen werden (**Tab. 1B**).

Mittels Pool-WGS konnten für alle Haplotypen die höchsten Korrelationen zur Individual-Genotypisierung erzielt werden. Dazu zeichneten sich die biologischen Wiederholungen noch durch eine geringe Abweichung zu einander aus. Da die Lesetiefe in den WGS-Haplotypen allesamt hoch waren (> 950x), wollten wir in einer Simulation testen, ab welchen Lesetiefen eine valide Aussage über die Allelfrequenz zu machen ist. Die Simulation zeigte, dass bereits ab einer Lesetiefe von 500x eine hohe Genauigkeit (Pearson Korrelation > 0,95) erzielt wird, mit Abstrichen auch schon ab 200x.

Fallbeispiel – Gerstenpopulation

Weiterführend haben wir die untersuchte BC_2F_{23}-Gerstenpopulation genutzt, um genomweite Änderungen der Allelfrequenz

Glossar

Allel (Kurzform von „allelomorph"): Ein Allel ist eine Variation derselben Nukleotidsequenz an derselben Stelle auf einem DNA-Molekül; der Ort eines Gens bzw. eines anderen genetischen Elements wird als Locus bezeichnet, und alternative DNA-Sequenzen an einem Locus werden Allele genannt.

Autopolyploidie (von auto-, Polyploidie): Polyploidie bezeichnet das Phänomen, das manche Arten mehr als zwei Sätze von Chromosomen in den Zellen besitzen. Bei der Autopolyploidie liegen die arteigenen Chromosomensätze vervielfältigt vor, während bei der Allopolyploidie jeweils artfremde Chromosomensätze die zwei oder mehr Sätze von Chromosomen bilden.

Contig: Ein Contig (von *contiguous*) ist ein Satz überlappender DNA-Segmente, die zusammen eine Konsensusregion der DNA darstellen.

GH (Gen-annotierter Haplotyp): Nukleotidsequenz eines Genes, zusammen gesetzt aus mehreren SNPs.

Haplotyp: Ein Haplotyp ist eine Gruppe von benachbarten Allelen in einem Organismus, die gemeinsam von einem einzigen Elternteil vererbt werden.

Marker: Ein genetischer Marker ist ein Gen oder eine DNA-Sequenz (z. B. SNP) mit einer bekannten Position auf einem Chromosom, die zur Identifizierung von Allelen, Individuen oder Arten verwendet werden kann.

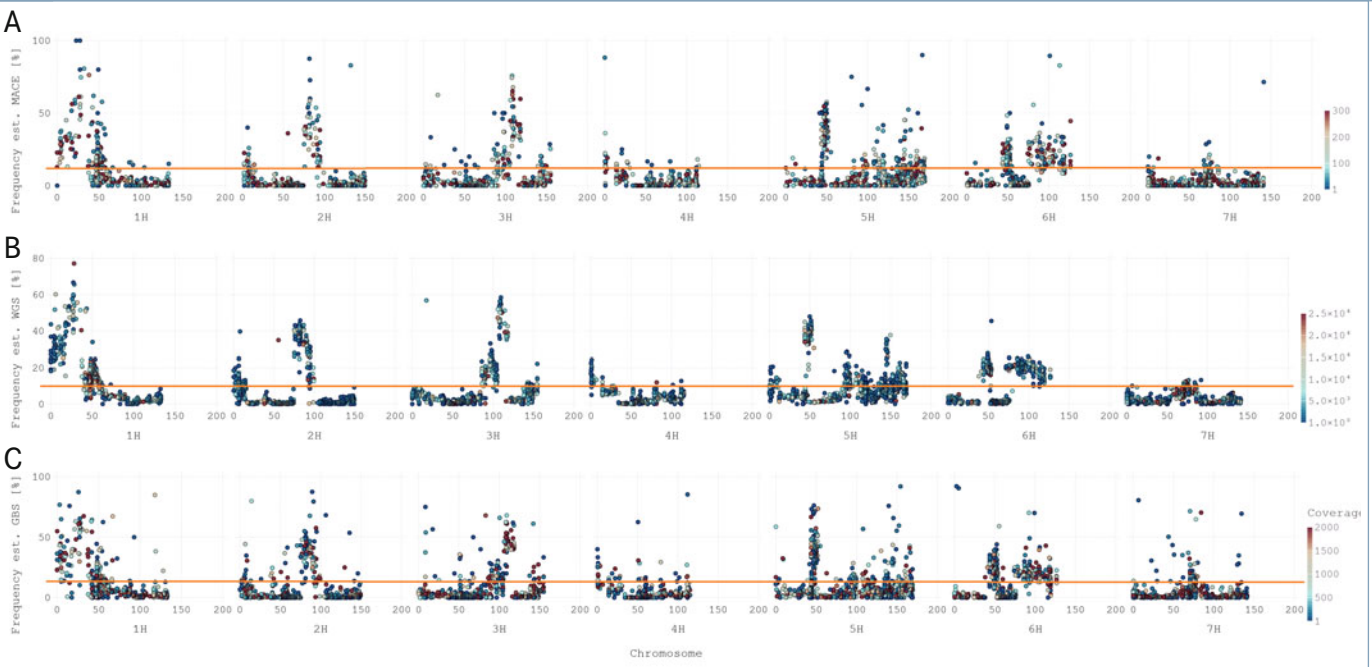
MH (Marker-annotierter Haplotyp): Nukleotidsequenz um einen Marker (SNP) herum.

Selective sweep: Ein vorteilhaftes Allel hat seine Häufigkeit in der Population stark erhöht oder liegt in der Population sogar fixiert vor (d. h. es wird eine Häufigkeit von 1 erreicht). Dieses führt zu einer Verringerung oder Eliminierung der genetischen Variation von benachbarten Nukleotidsequenzen, da Kopplungsbrüche relativ selten sind.

SNP (Einzelnukleotid-Polymorphismus): eine Substitution eines einzelnen Nukleotids an einer bestimmten Position im Genom.

Tab. 1: Pool-Genotypisierung: Validierung der abgeleiteten Allelfrequenzen. **A**, Pearson-Korrelationskoeffizienten der Individual- und Pool-Genotypisierungen auf verschiedenen Haplotyp-Ebenen zueinander (SNP: Einzel-Locus; GH: Gen-annotierter Haplotyp; MH: Marker-annotierter Haplotyp, beispielsweise von einem bekannten QTL; Contig: Genomabschnitt von definierter Länge). Der fehlende Wert der GBS-SNP-Level ist auf zu wenige direkte SNP-Übereinstimmungen zwischen KASP- und GBS-Pool-Seq zurückzuführen. **B**, Gleichheit (p-Werte) der 2–3-fach wiederholten Pool-Proben, gemessen mittels eines negativ binomialen Modells.

Wiederholungsvergleich	Haplotyp	GBS	MACE	WGS
A, Pearson-Korrelation (r) & Lesetiefe je Locus/Haplotyp (Anzahl)	SNP-Ebene	–	0,79	0,93*
	Lesetiefe	28	25	9
	GH-Ebene	0	0,9	0,97
	Lesetiefe	137	42	963
	MH-Ebene	0,83	0,93	0,96
	Lesetiefe	450	70	5,443
	Contig-Ebene	0,94	0,88	0,95
	Lesetiefe	4,837	520	74,855
B, negativ binomiales Model (p-Wert)	SNP-Ebene	< 0,001	< 0,001	< 0,001
	GH-Ebene	< 0,001	0,007	0,67
	MH-Ebene	0,17	0,35	0,52
	Contig-Ebene	0,91	0,46	0,35
	* Niedrige Anzahl der zu vergleichenden Marker, weshalb der Wert überschätzt wird.			



▲ **Abb. 3:** Genomweite Allelfrequenzmuster auf einer genetischen Karte, annotiert an Marker-Haplotypen (MH). Dargestellt sind die 7 Gersten Chromosomen (x-Achse) gegen die Allelfrequenz der eingekreuzten Wildform (y-Achse). Die Farbe der Punkte gibt Aufschluss über die Lesetiefe des Haplotyps. Die orange Linie zeigt die erwartete Wildform-Allelfrequenz in der BC_2F_1. **A,** MACE. **B,** WGS. **C,** GBS.

zu identifizieren (**Abb. 3**). Alle drei Methoden eignen sich, die größeren Allelfrequenzänderungen auf den Chromosomen 1H, 2H, 3H, 5H und 6H im Vergleich zur erwarteten Allelfrequenz der Ausgangspopulation (orange Linie) aufzudecken. WGS sorgt dabei für das geringste Hintergrundrauschen und die höchste Auflösung. ■

Literatur

[1] Gautier M, Foucaud J, Gharbi K et al. (2013) Estimation of population allele frequencies from next-generation sequencing data: Pool-versus individual-based genotyping. Mol Ecol 22: 3766–3779
[2] Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. Nat Rev Genet 15: 749–763
[3] Dreissig S, Mascher M, Heckmann S, Purugganan M (2019) Variation in Recombination Rate Is Shaped by Domestication and Environmental Conditions in Barley. Mol Biol Evol 36: 2029–2039
[4] Tilk S, Bergland A, Goodman A et al. (2019) Accurate allele frequencies from ultra-low coverage Pool-seq samples in evolve-and-resequence experiments. G3: Genes, Genomes, Genetics 9: 4159–4168

Funding note: Open Access funding enabled and organized by Projekt DEAL.

Korrespondenzadresse:
Prof. Dr. Jens Léon
Rheinische Friedrich-Wilhelms-Universität Bonn
Institut für Nutzpflanzenwissenschaften und Ressourcenschutz
Professur Pflanzenzüchtung
Katzenburgweg 5
D-53115 Bonn
j.leon@uni-bonn.de

AUTOREN

Jens Léon
Studium der Agrarwissenschaften an der Universität zu Kiel mit anschließender Promotion **1985** in Pflanzenzüchtung unter Anleitung von Prof. Dr. M. Hühn, Habilitation daselbst im Jahr **1992**. Bis **1996** Hochschulassistent an der Universität zu Kiel. Seit **1996** Professor für Pflanzenzüchtung an der Universität Bonn.

Agim Ballvora
Studium der Biologie. Wissenschaftlicher Mitarbeiter in der Pflanzengenetik am Max-Planck-Institut für Pflanzenzüchtungsforschung, Köln. **1995** Promotion. Anschließend Postdoktorand am Institut des Sciences Vegetale, CNRS, Frankreich und am Max-Planck-Institut für Pflanzenzüchtungsforschung. Seit **2010** Wissenschaftlicher Mitarbeiter beim INRES-Pflanzenzüchtung an der Universität Bonn.

Michael Schneider
2011–2016 Studium der Agrar- und Nutzpflanzenwissenschaften an der Universität Bonn. Dort bis **2020** Promotion in der Pflanzenzüchtung unter Anleitung von Prof. Dr. J. Léon. **2020–2022** Postdoktorand an der Universität Düsseldorf in der quantitativen Genetik bei Prof Dr. B. Stich. Folgend angestellt am Forschungsinstitut für biologischen Landbau (FiBL) im Department Nutzpflanzenwissenschaften, mit Fokus auf Züchtung für den organischen Anbau.