Genomanalyse der Wilden Weinrebe *Vitis vinifera* subsp. *sylvestris*

Dissertation

zur Erlangung des Grades

Doktor der Naturwissenschaften

(Dr. rer. nat.)

am Fachbereich Biologie der

Johannes Gutenberg-Universität Mainz

Institut für Molekulargenetik,

Gentechnische Sicherheitsforschung und -beratung

vorgelegt von

Sabine Fischer

Geboren am 26. September 1988

in Speyer

Mainz, 2017

Dekan:

1. Berichterstatter:

2. Berichterstatter:

Tag der mündlichen Prüfung: 21.07.2017

Inhaltsverzeichnis

1 Einleitung

1.1 Die Weinrebe, *Vitis vinifera* L.

Die Weinrebe zählt zu den wirtschaftlich und kulturell bedeutendsten Kulturpflanzen. Die Weintraube ist mit 77,2 Mio. Tonnen nach Wassermelonen, Bananen und Äpfeln die meist geerntete Obstart weltweit (FAO 2013). Über die Hälfte der Weintrauben wird zu Wein verarbeitet, die Nachfrage ist mit 240 Mio. Hektolitern jährlich ungebrochen hoch (OIV 2015). Daneben wächst aber auch der Anteil anderer Erzeugnisse der Weintraube. Sie findet vermehrt in Form von Tafeltrauben und Rosinen sowie in Säften und Kosmetika Verwendung. Weltweit wird der Genuss von Wein als Kulturgut geschätzt, ihm wird sogar, in Maßen konsumiert, eine gesundheitsfördernde Wirkung zugeschrieben. So senken Resveratrol und andere phenolische Verbindungen, die insbesondere in Rotweinen enthalten sind, das Risiko von Herz-Kreislauf-Erkrankungen (German & Walzem 2000). In weiteren Studien konnten – jedoch bisher meist nur *in vitro* – positive Effekte bei Krebs (Zhou et al. 2005), Diabetes (Brasnyó et al. 2011) und neurodegenerativen Erkrankungen, wie Alzheimer oder Parkinson (Tellone et al. 2015), nachgewiesen werden.

Obwohl die Gattung der Weinreben (*Vitis*) rund 60 Arten beinhaltet (Mullins et al. 1992), konzentriert sich die wirtschaftliche Nutzung fast ausschließlich auf die einzige in Europa einheimische Spezies *Vitis vinifera* L. Diese umfasst mit der Edlen Weinrebe *Vitis vinifera* subsp. *vinifera* und der Wilden Weinrebe *Vitis vinifera* subsp. *sylvestris* zwei Formen, die aufgrund ihrer morphologischen und physiologischen Unterschiede (Tabelle 1) gemeinhin als zwei Subspezies betrachtet werden (Zohary et al. 2012). Diese Unterteilung begründet sich jedoch nicht in einer geographischen Trennung der Subspezies, sie ist vermutlich vielmehr ein Produkt der Domestizierung der Weinrebe durch den Menschen (This et al. 2006). Die Wilde Weinrebe gilt als natürlicher Vorfahr der Edlen Weinrebe, die durch Kultivierung aus ersterer hervorgegangen ist (Zohary et al. 2012). Die primäre Domestizierung der Weinrebe fand vermutlich im Neolithikum im Nahen Osten statt. Dort wurden bei Ausgrabungen im Zāgros-Gebirge im heutigen Iran Tongefäße mit chemischen Rückständen entdeckt, die auf eine Produktion von Wein bereits 5400–5000 v. Chr. hindeuten (McGovern et al. 1996). *Simple Sequence Repeat* (SSR)-Analysen zahlreicher aktuellerer molekularbiologischer Studien liefern jedoch Hinweise auf weitere sekundäre Domestizierungsereignisse im Mittelmeerraum (De Andrés et al. 2012; Lopes et al. 2009; De

Mattia et al. 2008; Arroyo-García et al. 2006). Parallel dazu wird eine mögliche Introgression genetischen Materials lokaler Wildreben in das Genom moderner westeuropäischer Kulturreben diskutiert (Myles et al. 2011).

Folgen der Domestizierung sind zahlreiche morphologische und physiologische Veränderungen, die eine wirtschaftliche Nutzung vereinfachen und verbessern. Eine Übersicht hierzu liefert Tabelle 1.

Tabelle 1: Morphologische und physiologische Unterschiede zwischen *Vitis vinifera* subsp. *sylvestris* und *Vitis vinifera* subsp. *vinifera*

	Vitis vinifera* subsp. *sylvestris **Wilde Weinrebe**	***Vitis vinifera* subsp. *vinifera*** **Edle Weinrebe**
Blütenmorphologie	diözisch	hermaphroditisch
Bestäubung	Fremdbestäubung	Selbstbestäubung
Beeren	klein, rund oder abgeflacht, blauschwarz, säuerlich	groß, gestreckt, große Farbvielfalt, süßlich
Samen	klein, rund, großes Breiten/ Längen-Verhältnis (>0,7)	groß, birnenförmig, kleineres Breiten/ Längen-Verhältnis (<0,6)
Traubenbündel	klein, rund bis kegelförmig, Beerenreife nicht uniform	groß, verlängert, kompakt, uniforme Beerenreife
Blätter	klein, dreilappig	groß, ganzrandig oder flach gebuchtet
Vermehrung	sexuell und vegetativ	vegetativ, meist Pfropfung
Habitat	feuchte Böden	trockene Böden
Verbreitung	isolierte kleine Populationen entlang des mediterranen Beckens, Rhein- und Donauufer	weltweiter Anbau einiger weniger dominanter Kultursorten

Die Tabelle wurde auf Basis der Daten von Zohary et al. (2012), Zecca et al. (2010), This et al. (2006), Grassi et al. (2003, 2006) und Olmo (1976) erstellt.

Besonders bemerkenswert ist der Wandel der Blütenmorphologie. Bei weiblichen Pflanzen der diözischen Wilden Weinrebe kommt es zu einem frühen Abbruch bei der Staminaentwicklung, männliche Pflanzen hingegen bilden keine funktionellen Fruchtknoten aus und tragen infolgedessen keine Früchte (Caporali et al. 2003). Im Gegensatz dazu besitzt die Edle Weinrebe hermaphroditische Blüten mit voll entwickelten Stamina und Fruchtknoten. Die Selektion hermaphroditischer Pflanzen, die einen platzsparenden Anbau ermöglichen (alle Pflanzen tragen Früchte), war demnach ein wichtiger Schritt während der Domestizierung der Weinrebe (Grassi et al. 2003). Die der Entwicklung hermaphroditischer Blüten zugrundeliegenden molekularen Mechanismen sind bisher ungeklärt und daher ein zentrales Thema aktueller Forschung (Ramos et al. 2014; Picq et al. 2014; Fechter et al. 2012).

Die morphologischen Unterschiede können genutzt werden, um zwischen beiden Subspezies zu differenzieren. Da die Grenzen zwischen den Merkmalen jedoch oft diffus und Hybridisierungen zwischen den Subspezies möglich sind, wird in der Forschung zur zweifelsfreien Unterscheidung zwischen *Vitis vinifera* subsp. *vinifera* und *Vitis vinifera* subsp. *sylvestris* auf SSR-Marker zurückgegriffen. Die hohe Auflösung dieser molekularen Marker ermöglicht nicht nur eine Abgrenzung der beiden Subspezies voneinander, sie gestattet es auch, innerhalb der Kulturreben zwischen Rebsorten zu differenzieren sowie Verwandtschaftsverhältnisse aufzuklären (This et al. 2004; Bowers et al. 1999). Bei der Wilden Weinrebe kann mittels SSR-Markern sogar zwischen Populationen unterschieden werden (Grassi et al. 2008).

Die Wilde Weinrebe ist europaweit vom Aussterben bedroht. Hauptgrund hierfür ist die anhaltende, massive Zerstörung und Fragmentierung der natürlichen Habitate der als rankenden Liane wachsenden Wilden Weinrebe (Arnold et al. 1998). Meist ist es der Mensch, der beispielsweise durch Holzschlag, Flussbegradigungen, Waldbrände oder Flurbereinigungen in den Lebensraum der Wilden Weinrebe eingreift (Arnold et al. 2005). Hinzu kam in der Mitte des 19. Jahrhunderts die Ausbreitung nordamerikanischer Pathogene wie der Reblaus (*Phylloxera*) oder dem Echten Mehltau (*Oidium*) und dem Falschen Mehltau (*Plasmopara viticola*) in Europa (Levadoux 1956).

Heute existieren infolgedessen nur noch vereinzelte Populationen in alluvialen und kolluvialen Wäldern entlang des mediterranen Beckens und isolierte Randpopulationen z.B. in den Auenwäldern von Rheinland-Pfalz und Baden-Württemberg (Ledesma-Krist et al. 2015; Arnold et al. 1998). Das Verbreitungsgebiet reicht dabei, wie in Abbildung 1 gezeigt, von der südlichen Atlantikküste der iberischen Halbinsel bis zum westlichen Rand des Himalayas (Hegi 1925). Dass gerade die Überschwemmungsgebiete der Auenwälder einen Rückzugsort für die Wilde Weinrebe darstellen, ist nicht verwunderlich, da die wurzelbefallende Reblaus eine Überflutung nicht überlebt (Ocete & Lara 1994). Durch Eingriffe des Menschen in die Flussläufe sinkt jedoch der Wasserspiegel in den Auenwäldern und die Reblaus dringt weiter vor (Arroyo-García & Revilla 2013; Arnold et al. 2005).

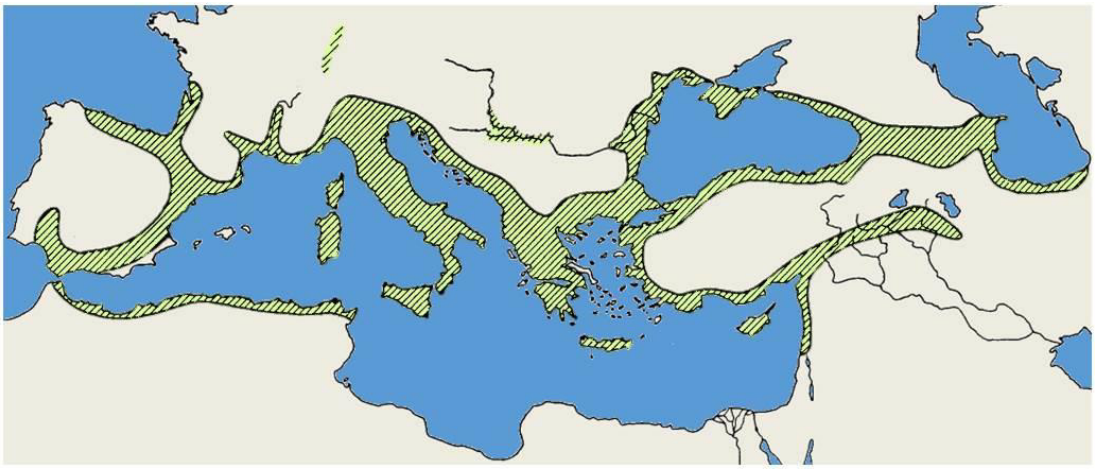

Abbildung 1: Verbreitung der Wilden Weinrebe *Vitis vinifera* subsp. *sylvestris*
Die Karte zeigt in grün die Verbreitungsgebiete der Wilden Weinrebe entlang des mediterranen Beckens in Europa, Nordafrika und Asien. Jenseits der östlichen Grenzen des gezeigten Ausschnitts existieren vereinzelte Populationen in Turkmenistan, Tadschikistan und Kasachstan. Die Abbildung wurde in Anlehnung an Zohary et al. (2012) erstellt.

Die vorhandenen Wildrebenpopulationen sind in der Regel recht klein und umfassen oft weniger als zehn Individuen (Arnold 2002). Diese geringe Individuenzahl erschwert in Kombination mit der Diözie und der geringen Migrationsrate der verhältnismäßig schweren Pollenkörner der Wilden Weinrebe die Reproduktion und somit ein Wachstum der Populationen (Di Vecchi-Staraz et al. 2009). Die meisten Wildrebenpopulationen finden sich in Spanien, Italien und dem Kaukasus (Biagini et al. 2014; Grassi et al. 2006). Diese Gebiete dienten während der letzten Eiszeit als Refugien für zahlreiche Pflanzenspezies, darunter auch die Wilde Weinrebe (De Mattia et al. 2008). Ausgehend von diesen Refugien fand inter- und postglazial eine rasche Rekolonialisierung Richtung Norden bzw. Nordwesten statt. Daher weisen diese südlichen Populationen eine höhere genetische Diversität als Wildreben in nördlicheren Gebieten auf, ein Phänomen, das als *southern richness* bezeichnet wird (Biagini et al. 2014; Grassi et al. 2008). Im Allgemeinen lässt sich dennoch beobachten, dass die Diversität der Wilden Weinrebe infolge ihrer kleinen und fragmentierten Populationen stark abgenommen hat und inzwischen sogar teilweise von der Edlen Weinrebe übertroffen wird (De Andrés et al. 2012; Barnaud et al. 2010). Letztere zeigt eine verhältnismäßig hohe Diversität, die mit der der Pappel vergleichbar ist (Smulders et al. 2008) und die der Tomate sogar übersteigt (Ranc et al. 2008). Begründet ist dies in der langen Kultivierungsgeschichte der Edlen Weinrebe, die durch sexuelle Kreuzungen gefolgt von anschließender vegetativer Vermehrung mit somatischen Mutationen charakterisiert ist (Pelsy 2010). Allerdings ist ein Großteil dieser genetischen Diversität lediglich auf

Sammlungen und Genbanken beschränkt, da im wirtschaftlichen Sektor die Monokultur einiger weniger Rebsorten dominiert (This et al. 2006). Diese genetische Erosion erhöht, sowohl auf Seite der Wilden Weinrebe als auch auf Seite der Kulturrebe, die Anfälligkeit gegenüber neuen Pathogenen und verringert gleichermaßen die Anpassungsfähigkeit an neue Umweltbedingungen (Arroyo-García & Revilla 2013; Arnold et al. 1998).

Neben dem allgemeinen Artenschutz zur Erhaltung der Biodiversität kommt dem Schutz der Wilden Weinrebe noch eine weitere besondere Bedeutung zu. Als nahe Verwandte der Kulturrebe stellt die Wilde Weinrebe *Vitis vinifera* subsp. *sylvestris* eine wichtige genetische Ressource für die Rebenzüchtung dar (Ledesma-Krist et al. 2015; Negrul 1938). Zu den vorteilhaften Eigenschaften der Wilden Weinrebe, die für die Züchter von Nutzen sein könnten, zählen unter anderem Resistenzen gegen verschiedene Pathogene und Erkrankungen, wie beispielsweise dem Echten und dem Falschen Mehltau sowie der Schwarzfäule (Schröder et al. 2015; Nick 2012; Ocete et al. 2008).

1.2 Crop Wild Relatives

Die Anforderungen an Kulturpflanzen verändern sich stetig. Mit dem Wachstum der Weltbevölkerung steigt die Nachfrage nach höheren Erträgen und verbessertem Nährstoffgehalt. Gleichzeitig fordert der Klimawandel Anpassung an neue abiotische Umweltfaktoren (Brozynska et al. 2015). Hinzu kommen neue Krankheitserreger, die altbewährte Pathogenresistenzen und Pflanzenschutzmittel entwaffnen (Anderson et al. 2004). Diese Herausforderungen bedürfen einer kontinuierlichen Züchtung und Verbesserung der Nutzpflanzen. Hierfür werden als Quelle neuer nützlicher Eigenschaften pflanzliche genetische Ressourcen (PGR) benötigt (Esquinas-Alcázar 1993). Eine solche Ressource stellen die *crop wild relatives* (CWRs) dar. Dabei handelt es sich um wilde Pflanzenspezies, die mehr oder weniger eng mit wirtschaftlich genutzten Kulturpflanzen verwandt sind (Maxted et al. 2006). Im Gegensatz zu Kulturpflanzen – bei denen es bedingt durch die Domestizierung zu einem Flaschenhalseffekt und infolgedessen zu einer genetischen Verarmung kam – beinhalten CWRs in der Regel eine große genetische Diversität, die von Züchtern genutzt werden kann (Vincent et al. 2013). Dieses große Potenzial der Wildpflanzen erkannte bereits der russische Pflanzengenetiker Vavilov Anfang des 20. Jahrhunderts (Vavilov 1938).

Seit den 1940er Jahren werden CWRs daher gezielt in Züchtungsprogrammen zur Verbesserung von Nutzpflanzen eingesetzt (Meilleur & Hodgkin 2004). Klassischerweise werden dabei vorteilhafte Eigenschaften der wilden Verwandten durch Introgression auf die Kulturpflanzen übertragen. Dabei liegt der Schwerpunkt klar auf dem Transfer von Resistenzeigenschaften; schätzungsweise 80 % der genutzten CWR-Gene vermitteln eine Abwehr von Pathogenen und Krankheiten (Hajjar & Hodgkin 2007). Daneben finden sich auch andere vorteilhafte Merkmale, wie eine verbesserte Toleranz gegenüber abiotischen Stressfaktoren, eine Steigerung des Ertrags oder der Fruchtqualität, die Verbesserung des Nährstoffgehalts und cytoplasmatische männliche Sterilität (Dwivedi et al. 2008). Einige Beispiele sind in Tabelle 2 gezeigt. Am besten gelingt der Gentransfer, wenn der Donor-CWR möglichst nah mit der Kulturpflanze verwandt oder idealerweise sogar ihr direkter Vorfahr ist (Maxted & Kell 2009). Dennoch haben Züchter mit Kreuzungsinkompatibilität, Reproduktionsbarrieren, sterilen Hybriden und dem sogenannten *linkage drag*, also dem Mitübertragen von unerwünschten negativen Eigenschaften der CWRs durch genetische Kopplung, zu kämpfen (Stebbins 1958; Zeven et al. 1983). Letzteres erfordert zeitintensive Rückkreuzungen mit anschließender Selektion, um die unwillkommenen nachteiligen Gene wieder zu beseitigen (Hajjar & Hodgkin 2007). Neue molekulargenetische Verfahren können hilfreich sein, den Einsatz von CWRs in der Züchtung zu erleichtern und zu beschleunigen. Hierbei kann auf verschiedenen Ebenen in den Züchtungsprozess eingegriffen werden. Den Ausgangspunkt für die Verbesserung der Nutzpflanzen bildet stets die Identifizierung nützlicher oder erwünschter Gene im Genom der CWRs. *Next-Generation Sequencing* ermöglicht eine schnelle und kostengünstige Genom- und Transkriptomanalyse (Brozynska et al. 2015). Ohne Vorabwissen kann genomweit nach Bereichen erhöhter oder erniedrigter Diversität gesucht und selektierte Varianten identifiziert werden. Falls hingegen bereits Kandidatengene bekannt sind, gestattet das *targeted resequencing* die Identifizierung zugrundeliegender Mutationen im Hochdurchsatzmaßstab (Ford-Lloyd et al. 2011). Techniken wie *embryo rescue* ermöglichen es, zwischenartliche Kreuzungsbarrieren zu überwinden und auf diese Weise Hybridisierungen über Art-, Familien- und sogar Gattungsgrenzen hinweg durchzuführen (Kaneko & Bang 2014). Im Anschluss an die Kreuzung und im Verlauf von Rückkreuzungen gestatten molekulare Marker eine Überwachung der Introgression und die Selektion gewünschter Genotypen (Francia et al. 2005).

Tabelle 2: Ausgewählte Beispiele für CWRs als genetische Ressource in der Züchtung

Nutzen	Kulturpflanze	CWR	Eigenschaft	Quelle
Pathogen- bzw. Krankheitsresistenz	Tomate (*Solanum lycopersicum*)	u.a. *Solanum peruvianum*, *Solanum cheesmaniae*, *Solanum pennellii*	Resistenzen gegen über 40 Pathogenen	Rick & Chetelat 1995
	Kulturapfel (*Malus domestica)*	*Malus floribunda*	Resistenz gegen Apfelschorf	Brown 1975
	Edle Weinrebe (*Vitis vinifera*)	*Vitis berlandieri*, *Vitis riparia*, *Vitis rupestris*	Resistenz gegen die Reblaus	This et al. 2006
	Mais (*Zea mays*)	*Zea diploperennis, Zea perennis*	Resistenz gegen diverse Viren	Nault et al. 1982
	Gerste (*Hordeum vulgare*)	*Hordeum bulbosum*	Resistenz gegen Echten Mehltau	Xu & Kasha 1992
	Mohrenhirse (*Sorghum bicolor*)	*Sorghum macrospermum*	Resistenz gegen die Mücke *Stenodiplosis sorghicola*	Price et al. 2005
Toleranz gegenüber abiotischem Stress	Weizen (*Triticum aestivum*)	*Thinopyrum bessarabicum*	Salztoleranz	King et al. 1997
	Reis (*Oryza sativa*)	*Oryza rufipogon*	Trockenheitstoleranz	Zhang et al. 2006
	Reis (*Oryza sativa*)	*Oryza rufipogon*	Aluminiumtoleranz	Nguyen et al. 2003
	Edle Weinrebe (*Vitis vinifera*)	*Vitis amurensis*	Kältetoleranz	He et al. 1981
	Kulturapfel (*Malus domestica)*	*Malus baccata*	Kältetoleranz	Cummins et al. 1979
Verbesserung des Nährstoffgehalts	Soja (Glycine max)	*Glycine soja*	Proteingehalt	Sebolt et al. 2000
	Weizen (*Triticum aestivum*)	*Triticum turgidum*, *Aegilops tauschii*	Zink- und Eisengehalt	Ogbonnaya et al. 2013
Ertragssteigerung	Raps (*Brassica napus*)	*Brassica rapa*, *Brassica oleracea*	Mehr Samen	Osborn et al. 2007
Männliche Sterilität	Straucherbse (*Cajanus cajan*)	*Cajanus* cajanifolius	Cytoplasmatische männliche Sterilität	Saxena et al. 2005

Die Tabelle wurde in Anlehnung an Maxted und Kell (2009), Dwivedi et al. (2008) sowie Hajjar und Hodgkin (2007) zusammengestellt.

Trotz ihres Potenzials in der Züchtung ist die Biodiversität der CWRs bedroht. Ursache ist zumeist eine Zerstörung des Lebensraums durch den Menschen (Bilz et al. 2011). Schätzungsweise sind mehr als 16 % der europäischen CWR-Spezies aktuell oder in naher Zukunft gefährdet bzw. vom Aussterben bedroht (Maxted & Kell 2012). Ungeachtet ihres Nutzens für die Zukunft der Landwirtschaft werden CWRs in Naturschutzprogrammen bisher vernachlässigt. Nur etwa 6 % der europäischen CWR-Spezies sind Teil von Genbanken oder vergleichbaren *ex situ*-Sammlungen (Ford-Lloyd et al. 2011). *In situ*-Programme, die gezielt zum Schutz von CWR-Pflanzen eingerichtet wurden, sind selten. Doch gerade die *in situ*-Herangehensweise hat den Vorteil, dass die Pflanzen in freier Wildbahn Teil des Ökosystems bleiben und daher weiterhin der natürlichen Selektion unterliegen. Demnach stellen sie auf diese Weise eine dynamische Ressource dar, die sich wechselnden Umweltbedingungen anpasst (Brozynska et al. 2015). Um dies zu unterstützen, haben in den vergangenen Jahren verschiedene nationale und internationale Initiativen damit begonnen, CWRs in den Fokus des allgemeinen Bewusstseins zu rücken und Naturschutzprojekte ins Leben zu rufen (ITPGRFA 2009; CWRIS 2005). Zumeist wird als Erstes eine Inventarisierung und Priorisierung durchgeführt um zu klären, welche Spezies überhaupt als wilde Verwandte der einzelnen Kulturpflanzen in Frage kommen und ob sie eines besonderen Schutzes bedürfen (Heywood et al. 2007). Nicht immer sind die direkten Vorfahren der Kulturpflanzen bekannt und in manchen Fällen sind sie bereits ausgestorben. In anderen Fällen haben im Zuge einer Hybridisierung oder Polyploidisierung mehrere Spezies zum Genom der Kulturpflanzen beigetragen (Brozynska et al. 2015). Daher ist eine differenziertere Definition des Begriffs CWR notwendig. Harlan und De Wet (1971) etablierten hierzu das Genpool-Konzept (Abbildung 2). Auf Basis ihrer Fähigkeit zur Hybridisierung mit der Kulturpflanze lassen sich die verwandten Pflanzenspezies einem primären, sekundären oder tertiären Genpool zuordnen. Der tertiäre Genpool markiert hierbei die äußerste Grenze dessen, was tatsächlich noch als genetische Ressource in der Züchtung nutzbar ist. Jedoch ist eine Unterteilung in die einzelnen Genpools nur dann möglich, wenn ausreichend Informationen zu der Kreuzungsfähigkeit der einzelnen Spezies untereinander vorliegen. Dies trifft aber nicht auf alle Kulturpflanzen zu.

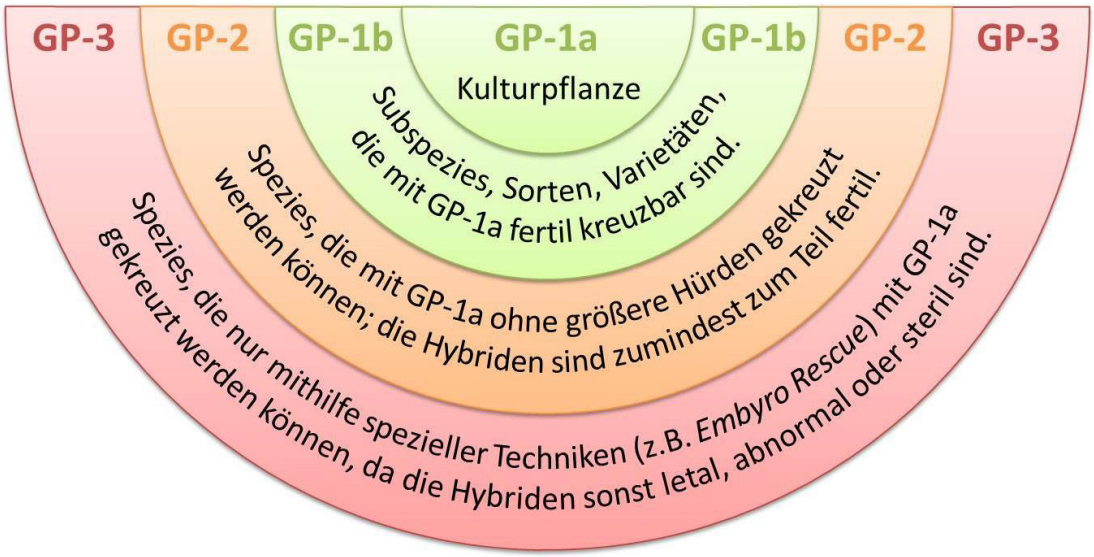

Abbildung 2: Genpool-Konzept
Die Abbildung zeigt schematisch das Genpool-Konzept nach Harlan und De Wet (1971). Kulturpflanzen und ihre Verwandten lassen sich einem primären (grün), sekundären (orange) und tertiären (rot) Genpool (GP) zuordnen. Grundlage für die Eingliederung in die verschiedenen GPs bildet die Fähigkeit zur Hybridisierung mit der Kulturpflanze. Die Grafik wurde in Anlehnung an die Publikation von Harlan und De Wet (1971) erstellt.

Daher entwarfen Maxted et al. (2006) als Ergänzung für Fälle, bei denen bisher keine Hybridisierungs-Experimente durchgeführt wurden, das Taxon-Konzept (Tabelle 3). Dieses orientiert sich an existierenden taxonomischen Hierarchien und beruht auf der Annahme, dass taxonomische Entfernung mit genetischer Distanz korreliert. Somit ist es praktisch auf alle Spezies anwendbar.

Tabelle 3: Taxon-Konzept

Taxongruppe (TG)	Mitglieder
TG-1a	Kulturpflanze
TG-1b	Gleiche Spezies wie die Kulturpflanze
TG-2	Gleiche Sektion/Serie wie die Kulturpflanze
TG-3	Gleiche Untergattung wie die Kulturpflanze
TG-4	Gleiche Gattung wie die Kulturpflanze
TG-5	Gleicher Tribus, aber andere Gattung als die Kulturpflanze

Gleich welches Konzept herangezogen wird, um die CWR-Spezies der jeweiligen Kulturpflanzen zu bestimmen, die von Naturschützern geforderte Konsequenz ist stets dieselbe: Ein systematischer Schutz der bedrohten Wildpflanzen sowohl *in situ* als auch *ex situ*. Nur auf diese Weise kann die für die Züchtung und Verbesserung von Nutzpflanzen

notwendige Diversität der pflanzlichen genetischen Ressourcen langfristig erhalten werden (Maxted & Kell 2009).

1.3 Fußspuren der Selektion

Die natürliche Selektion (lat. *selectio*, „Auslese“) ist ein durch Umweltfaktoren angetriebener Evolutionsprozess, der auf drei Arten auf Individuen und Populationen einwirken kann: positiv, negativ und balancierend (De Simoni Gouveia et al. 2014). Das Prinzip der positiven Selektion wurde bereits 1858 im sogenannten Ternate-Manuskript von Darwin und Wallace beschrieben. Neu entstandene vorteilhafte Mutationen werden positiv selektiert, da sie die Fitness ihres Trägers erhöhen (Sabeti et al. 2006). Infolgedessen steigt die Frequenz der entsprechenden Allele in der Population. Bei der negativen Selektion werden hingegen neue nachteilige Varianten aus der Population entfernt (Charlesworth et al. 1993). Daher wird sie auch als reinigende Selektion bezeichnet. Die balancierende Selektion steigert die Vielfalt in der Population, wie etwa in Fällen von sogenannter Heterosis, bei der Heterozygote einen Selektionsvorteil aufweisen (Charlesworth 2006). Im Gegensatz zur natürlichen Selektion fungiert bei der künstlichen Selektion nicht die Umwelt als Antriebskraft sondern der Mensch, der als Züchter tätig wird. Er selektiert Individuen mit ausgewählten Eigenschaften und schließt andere Individuen, die die gewünschten Merkmale nicht aufweisen, von der Fortpflanzung aus (Hammer 1984).

Jegliche Form der Selektion hinterlässt molekulare Fußspuren im Genom, da sie die Variationsstruktur der betroffenen genomischen Region verändert (Oleksyk et al. 2010). Anhand dieser Fußspuren, die auch Selektionssignale genannt werden, lassen sich Loci, die unter Selektion standen oder stehen, im Genom identifizieren (Qanbari et al. 2012). Die Möglichkeiten des *Next-Generation Sequencings* und die Verbesserung statistischer Methoden haben die genomweite Suche nach Fußspuren der Selektion in den vergangenen zehn Jahren deutlich vereinfacht und populär gemacht. Ziel zahlreicher Studien war und ist es, mehr über die grundlegenden evolutionären Prozesse zu lernen, die Genome formen und verändern (De Simoni Gouveia et al. 2014). Das gesteigerte Interesse an den Fußspuren der Selektion rührt daher, dass sie die Möglichkeit bieten, Rückschlüsse auf die Funktion von Genen und genomischen Regionen zu ziehen. Dem zugrunde liegt die Annahme, dass Positionen im Genom, die unter Selektion standen, eine funktionelle Bedeutung haben müssen (Nielsen 2005). So kann es gelingen, kausale Mutationen für bestimmte

Phänotypen, wie beispielsweise Erbkrankheiten des Menschen oder Veränderungen, die mit der Domestizierung von Nutztieren und Kulturpflanzen einhergehen, zu identifizieren. In Tabelle 4 sind hierfür einige ausgewählte Beispiele dargestellt.

Tabelle 4: Beispiele für Kandidatengene, die mittels genomweiter Suche nach Fußspuren der Selektion identifiziert wurden

Organismus	Gen	selektierte Funktion	Quelle
Mensch	*LCT*	Lactase-Expression nach der Jugend	Voight et al. 2006
Mensch	*TLR5*	Erkennung von bakteriellem Flagellin	Fagny et al. 2014
Schwein	*MAPK1*	Wachstum und Muskelentwicklung	Amaral et al. 2011
Hund	*AMY2B*	stärkereiche Ernährung	Axelsson et al. 2013
Huhn	*TSHR*	stetige, d.h. nicht-saisonale Fortpflanzung	Rubin et al. 2010
Darwinfinken	*HMGA2*	Schnabelgröße	Lamichhaney et al. 2016
Sojabohne	*FATB*	Ölgehalt der Bohne	Zhou et al. 2015
Mais	*MADS56*	Regulation der Blütezeit	Hufford et al. 2012

Wie die molekularen Fußspuren im Genom genau entstehen, lässt sich anhand des Beispiels der positiven Selektion verdeutlichen. Wenn eine neu entstandene vorteilhafte Mutation selektiert wird, steigt ihre Frequenz in der Population. Aufgrund der genetischen Kopplung verändert dieser Prozess nicht nur die Allelfrequenzen des betroffenen Locus, sondern auch die benachbarter Loci. Dieses Prinzip des „Mit-Selektierens“ gekoppelter, neutraler Nachbarallele wurde erstmals 1974 von Maynard Smith und Haigh beschrieben und wird seither als *hitchhiking effect* bezeichnet. Als Resultat entsteht, wie in Abbildung 3 gezeigt, ein sogenannter *selective sweep*.

Ein *selective sweep* geht mit vielfachen Veränderungen in der betroffenen Region einher (De Simoni Gouveia et al. 2014). Zu diesen zählt, dass Polymorphismen in der Umgebung eliminiert werden, wodurch es zu einer Abnahme der segregierenden Stellen (*number of segregating sites, S*) kommt (Kaplan et al. 1989). Ein weiteres Kennzeichen ist eine lokale Reduktion der durchschnittlichen Heterozygosität in der Region des *selective sweeps* (Maynard Smith & Haigh 1974). Der Überschuss an seltenen Varianten verursacht wiederum eine Verschiebung des Frequenzspektrums (*site frequency spectrum, SFS*) in der Umgebung des selektierten Abschnitts (Braverman et al. 1995). Weiterhin lassen sich ungewöhnlich weit ausgedehnte Haplotypen aufgrund einer Zunahme des Kopplungsungleichgewichts (*linkage disequilibrium, LD*) beobachten (Kim & Stephan 2002). Vergleicht man die kodierenden Sequenzen orthologer Gene, zeigt sich, dass im Falle der positiven Selektion die Zahl nicht-synonymer Austausche (d_N) signifikant größer ist als die Zahl synonymer Austausche (d_S) (Nei 2005).

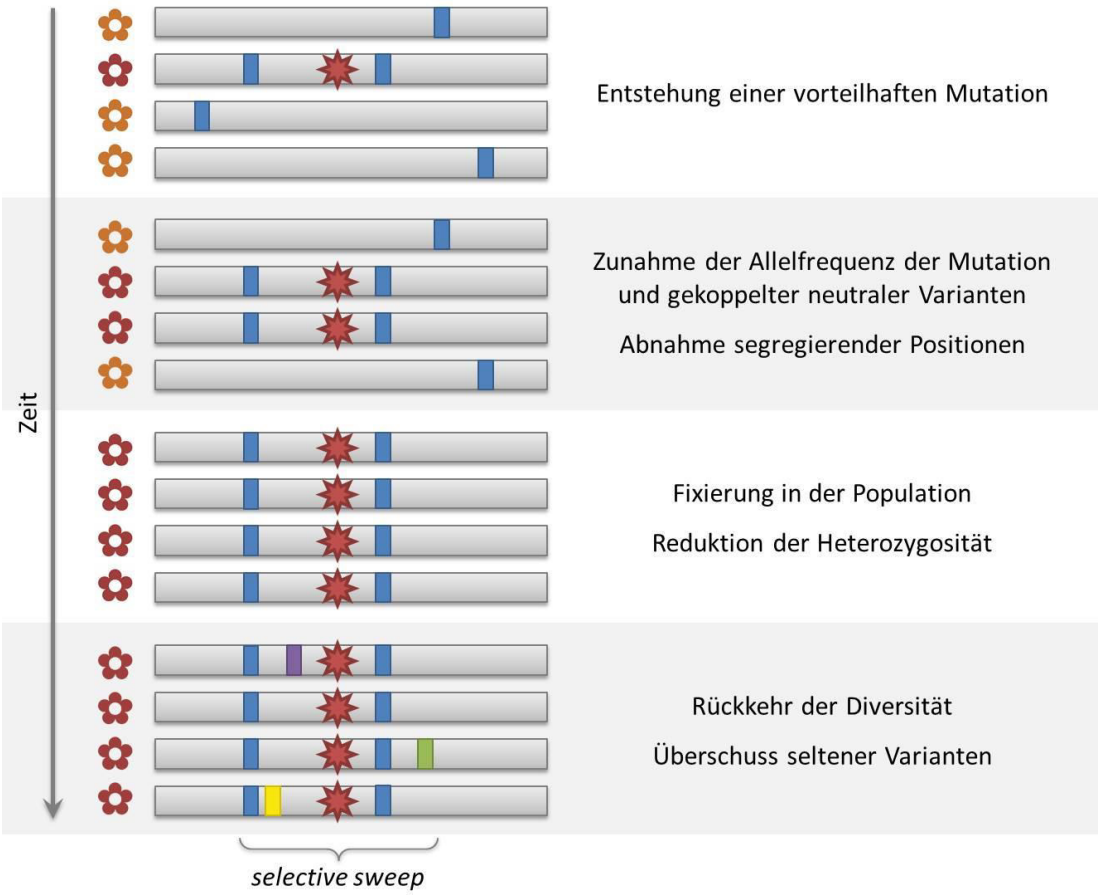

Abbildung 3: Entstehung eines *selective sweeps*
Dargestellt sind Veränderungen, die in einer genomischen Region stattfinden, wenn eine neu entstandene Mutation in einer Population positiv selektiert wird. Die Abbildung wurde in Anlehnung an die Publikation von Vitti et al. (2013) erstellt.

Jede einzelne dieser Veränderungen kann genutzt werden, um potenzielle *selective sweeps* im Genom aufzuspüren. Daher wurden in den vergangenen Jahren zahlreiche statistische Herangehensweisen entwickelt, um eine genomweite Suche nach Selektionssignalen durchzuführen. Auf eine detaillierte Beschreibung der einzelnen Methoden soll an dieser Stelle verzichtet und nur auf einige grundlegende Gemeinsamkeiten hingewiesen werden (siehe dazu Reviews von De Simoni Gouveia et al. (2014) und Oleksyk et al. (2010)). In der Regel haben die statistischen Tests eine der beschriebenen Veränderungen im Variationsmuster als Grundlage und vergleichen die beobachteten Werte mit den Erwartungen unter der Nullhypothese der selektiven Neutralität (Vitti et al. 2013; Kimura 1983). Typischerweise werden die Daten mithilfe eines *sliding windows* erhoben. Dabei handelt es sich um ein Fenster definierter Größe, das das Genom entlanggleitet und bei jedem Halt die für die jeweilige Methode spezifischen Werte für die einzelnen Positionen

innerhalb des im Fenster liegenden genomischen Abschnitts zusammenfasst (Carlson et al. 2005). Durch diese Vorgehensweise, denselben statistischen Test auf viele Positionen innerhalb einer genomischen Region gleichzeitig und additiv anzuwenden, wird die Falsch-Positiv-Rate signifikant erniedrigt (Vitti et al. 2013). Die Wahl der Methode hängt davon ab, welche Daten zur Verfügung stehen und wie alt die Fußspuren der Selektion sind, die identifiziert werden sollen. Der Vergleich homologer Sequenzen zwischen verschiedenen Spezies kann genutzt werden, um sehr alte Selektionsereignisse zu detektieren (Yang 2002). Hierbei liegt der Schwerpunkt somit primär auf makroevolutionären Prozessen. Populationsgenetische Daten erlauben hingegen einen Blick auf die Mikroevolution innerhalb einer Spezies. Die auf diese Weise identifizierbaren Fußspuren der Selektion sind dementsprechend jüngeren Datums wie zum Beispiel lokale Adaptionsprozesse im Anschluss an die *out-of-Africa*-Ausbreitung des *Homo sapiens* (Sabeti et al. 2006).

Problematisch wird die eindeutige Identifikation von Selektionssignalen, wenn innerhalb eines chromosomalen Abschnitts eine simultane oder zeitlich versetzte Interaktion zwischen verschiedenen Selektionsformen stattgefunden hat (Oleksyk et al. 2010). In solchen Fällen können die Einzelsignale miteinander verschwimmen oder sich gegenseitig aufheben. Hinzu kommt, dass in natürlichen Populationen allgegenwärtige demografische Prozesse wie Migration, Expansion, Aufgliederung und Flaschenhalseffekte Spuren im Genom hinterlassen, die Selektionssignalen gleichen und daher fälschlicherweise als solche interpretiert werden können. Diese Faktoren werden in den verwendeten statistischen Tests zum Teil nur unzureichend berücksichtigt oder erfordern komplexe mathematische Simulationen (Nielsen 2005; Wall et al. 2002). Hilfreich kann in all diesen Fällen eine Kombination der statistischen Methoden sein. Lassen sich in einer Region mit verschiedenen Herangehensweisen wiederholt Signale feststellen, liefert dies stichhaltigere Hinweise auf ein tatsächlich stattgefundenes Selektionsereignis und ermöglicht eine genauere räumliche Auflösung (Grossman et al. 2010). Letztendlich liefern diese *in silico*-Analysen jedoch immer nur Hinweise auf Kandidatengene, die potenziell unter Selektion standen. Eine endgültige Validierung dieser Kandidatengene bedarf der experimentellen Untersuchung ihrer Funktion im Phänotypen (Kamberov et al. 2013).

Nicht außer Acht gelassen werden sollte ferner die Tatsache, dass Selektion auch jenseits der Nukleotidebene wirkt. Neben der aufgrund ihrer einfachen Detektierbarkeit häufig im Fokus der Forschung stehenden Mutationen einzelner Basenpaare, können verschiedene

größere genetische Veränderungen ebenfalls Angriffspunkte der Selektion darstellen (Vitti et al. 2013). Hierzu zählen strukturelle Veränderungen des Erbguts, wie *copy number variations* (CNV), Mikrosatelliten und transposable Elemente sowie chromosomale Rearrangements, wie Inversionen und Translokationen. Nicht selten handelt es sich dabei um Ziele negativer Selektion, da solch schwerwiegende Veränderungen häufig nachteilige Effekte im Phänotypen hervorrufen (Cook & Scherer 2008). Jedoch sind auch Fälle bekannt, bei denen durch strukturelle Varianten ein Selektionsvorteil vermittelt wurde, beispielsweise im Falle des Amylase-Gens bei Säugern, dessen steigende Kopienzahl eine Anpassung an die stärkereiche Ernährung von Mensch und Hund darstellt (Perry et al. 2007; Axelsson et al. 2013). Jenseits der DNA-Ebene beeinflussen epigenetische Modifikationen u. a. den Verpackungsgrad des Erbguts und somit die Expression von Genen. Auch sie sind vererbbar und daher ein potenzieller Angriffspunkt für Selektion (Jablonka & Raz 2009). Seit einigen Jahren mehren sich die Hinweise, dass diese Form der Selektion gerade bei Evolutionsprozessen im Pflanzenreich eine große Rolle spielt (Hirsch et al. 2012; Richards 2011).

1.4 Architektur und Sequenzierung von Pflanzengenomen

Schätzungen zufolge bevölkern mehr als 400.000 verschiedene Spezies von Samenpflanzen die Ökosysteme der Erde (Govaerts 2001). Dabei zeigen sie eine enorme Formenvielfalt, die von der nur Millimeter kleinen Zwergwasserlinse bis hin zu über 100 Meter hohen Mammutbäumen reicht (Weber 2012). Diese Diversität spiegelt sich auch in den Genomen der Pflanzen wider. Die Genomgröße überspannt vier Größenordnungen und erstreckt sich dabei von knapp 60 Mb der carnivoren Reusenfallen (*Genlisea tuberosa* und *aurea*; Fleischmann et al. 2014) bis zu 152.000 Mb der Japanischen Einbeere (*Paris japonica*; Pellicer et al. 2010). Diese großen Unterschiede rühren insbesondere daher, dass sich in den Genomen zahlreicher Pflanzenspezies transposable Elemente angereichert haben, während andere Arten diese aktiv, beispielsweise durch illegitime Rekombination, aus ihrem Genom entfernen (Michael 2014). Daraus resultieren deutliche Unterschiede im prozentualen Anteil der repetitiven Elemente am Genom der Pflanzen (El Baidouri & Panaud 2013). Eine weitere Ursache für die Dynamik in der Genomgröße sind sogenannte *whole genome duplications* (WGD). Dabei handelt es sich um genomweite Duplikationen, auf die ausgeprägte chromosomale Rearrangements und ein Verlust redundanter Gene

folgen (Proost et al. 2011). Jedoch werden nicht alle verdoppelten Gene aus dem Genom entfernt oder stillgelegt. In einigen Fällen bleiben beide Kopien in ihrer ursprünglichen Form erhalten oder es kommt zu einer Neo- oder Subfunktionalisierung einer Genvariante (Adams & Wendel 2005). Da WGDs wiederholt in der Evolution der Angiospermen auftraten, ist ein Großteil der rezenten, offenkundig diploiden Pflanzenspezies tatsächlich paläopolyploid (Kellogg & Bennetzen 2004). Aber auch eine andauernde Polyploidie ist bei höheren Pflanzen weit verbreitet (Hegarty & Hiscock 2008). Gerade unter den Kulturpflanzen finden sich zahlreiche Spezies oder Varietäten – darunter Weizen, Kartoffeln und Bananen – mit vervielfachten Chromosomensätzen (Hilu 1993).

Eine Möglichkeit, Einblicke in die komplexe Architektur von Pflanzengenomen zu gewinnen, bieten seit Ende des 20. Jahrhunderts w*hole genome* Sequenzierungen. Aufgrund des Modellcharakters und der verhältnismäßig kleinen Genomgröße von 125 Mb wurde *Arabidopsis thaliana* als erster Vertreter der Pflanzenwelt für ein Genomprojekt ausgewählt (TAGI 2000). Hierbei wurde das Erbgut in durchschnittlich 100 kb große Bruchstücke zerlegt und kloniert. Die einzelnen Klone wurden in einer physikalischen Karte angeordnet und nach Sanger sequenziert (u. a. Mozo et al. 1998). Diese *clone-by-clone*-Strategie bietet den Vorteil, dass die Komplexität der Klone im Vergleich zu der des Gesamtgenoms deutlich reduziert ist und somit die Assemblierung der Sequenzdaten erleichtert wird (Green 1997). Das Resultat ist eines der wenigen fast vollständig assemblierten Genome mit nur vereinzelten Lücken, meist in Bereichen hochrepetitiver Regionen wie der Zentromere oder der rDNA. Nicht grundlos gilt das publizierte *Arabidopsis thaliana* Genom daher bis heute als eine Art Goldstandard für die Sequenzierung von Pflanzengenomen (Hamilton & Robin Buell 2012). Durch den hohen Kosten- und Zeitaufwand findet die *clone-by-clone*-Methode insbesondere bei größeren Pflanzengenomen selten Anwendung (Bolger et al. 2014). Daher wurde bei nachfolgenden Projekten, wie beispielsweise der Sequenzierung der Pappel oder der Weinrebe, auf *whole genome shotgun* (WGS)-Strategien zurückgegriffen (Tuskan et al. 2006; Jaillon et al. 2007). Bei dieser im Deutschen als Schrotschussverfahren bezeichneten Vorgehensweise wird das Genom in Form einer großen Anzahl zufällig generierter Fragmente direkt sequenziert (Weber & Myers 1997). Jedoch ist die Assemblierung der Chromosomen bei WGS-Sequenzierungen erschwert und erfordert hochentwickelte Computeralgorithmen (Imelfort & Edwards 2009). Trotz erheblicher Fortschritte in der Bioinformatik entsteht bei der Assemblierung der Sequenzen meist nur ein sogenanntes

draft genome, also ein Genom-Entwurf, der aus vielen Tausend *contigs* oder hunderten *scaffolds* mit zahlreichen kleineren und größeren Lücken besteht (Claros et al. 2012).

Die Entwicklung neuer *Next-Generation Sequencing* (NGS)-Technologien ermöglicht mit ihrem hohen Durchsatz und den vergleichsweise geringen Kosten Genomsequenzierungen nun auch kleineren Arbeitsgruppen. Dies zeigt sich im rasanten Anstieg der Publikationen von Pflanzengenomen in den vergangenen zehn Jahren (Abbildung 4).

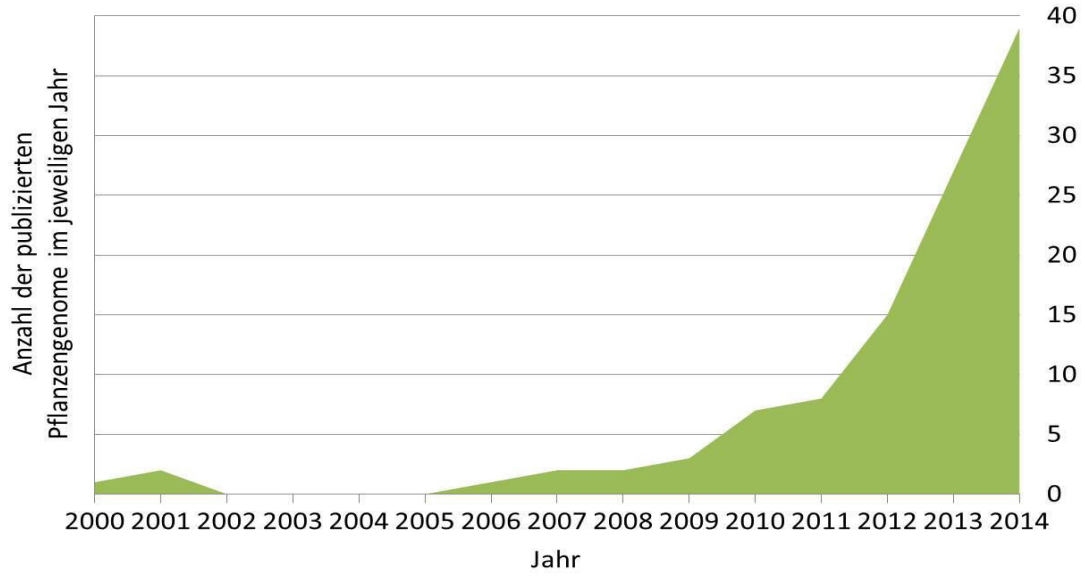

Abbildung 4: Anstieg der Publikationen von Pflanzengenomen
Das Diagramm zeigt die Anzahl publizierter Pflanzengenome pro Jahr im Zeitraum von 2000 bis 2014. Berücksichtigt wurden ausschließlich in Fachzeitschriften mit *peer-review*-Verfahren veröffentlichte Genome. Nicht einbezogen wurden Resequenzierungen von Populationen. Die verwendeten Zahlen stammen von PlaBi (http://www.plabipd.de/).

Bisher wurden Genomentwürfe von über 100 verschiedenen Pflanzenspezies veröffentlicht. Der Schwerpunkt liegt dabei klar auf wirtschaftlich genutzten Kulturpflanzen, sie stellen fast zwei Drittel der sequenzierten Pflanzengenome (Michael & VanBuren 2015). Aber auch das Interesse an wilden Verwandten und Modellorganismen mit außergewöhnlicher Genomarchitektur steigt (Henry 2012). Tabelle 5 vergleicht beispielhaft die Charakteristika einiger sequenzierter Pflanzengenome. Dabei fällt auf, dass auch nach der Etablierung der NGS-Technologien nicht vollständig auf die Sanger-Sequenzierung verzichtet wurde. Aktuell werden häufig Hybrid-Ansätze gewählt, bei denen verschiedene NGS-Plattformen und die traditionelle Sanger-Sequenzierung kombiniert werden (Hamilton & Robin Buell 2012).

Tabelle 5: Charakteristika ausgewählter sequenzierter Pflanzengenome

Spezies	Größe [Mb]	Genzahl	Rep. [%][1]	Sanger	454	Illumina	Quelle
Acker-Schmalwand (*Arabidopsis thaliana*)	125	25.498	14	×			TAGI 2000
Edle Weinrebe (*Vitis vinifera*)	475	30.434	41	×			Jaillon et al. 2007
Mais (*Zea mays*)	2300	32.540	85	×			Schnable et al. 2009
Kulturapfel (*Malus domestica*)	742	57.386	67	×	×		Velasco et al. 2010
Tomate (*Solanum lycopersicum*)	900	34.727	63	×	×		Sato et al. 2012
Zwerg-Wasserschlauch (*Utricularia gibba*)	77	28.500	3	×	×	×	Ibarra-Laclette et al. 2013
Weihrauch-Kiefer (*Pinus taeda*)	23.200	50.172	82			×	Neale et al. 2014
Weichweizen (*Triticum aestivum*)	17.000	124.201	81			×	Marcussen et al. 2014
Robusta-Kaffee (*Coffea canefora*)	710	25.574	50	×	×	×	Denoeud et al. 2014

[1] Rep. [%] = prozentualer Anteil repetitiver Elemente am Genom

Die kurzen Sequenzlängen (*reads*), die durch NGS-Technologien, wie z. B. die der Geräte der Firma Illumina, generiert werden, gepaart mit den im Vergleich zur Sanger-Sequenzierung höheren Fehlerraten, stellen die Forscher bei der Assemblierung der Pflanzengenome jedoch vor große Herausforderungen (Claros et al. 2012). Zusätzlich erschwert wird eine *de novo* Assemblierung der Sequenzdaten durch die bereits angesprochenen typischen Eigenschaften der Pflanzengenome. Computeralgorithmen sind nicht in der Lage, die mittel- und hochrepetitiven Bereiche, die einen Großteil der meisten Pflanzengenome ausmachen, korrekt zu assemblieren, sobald die Länge der repetitiven Einheiten die Länge der einzelnen *reads* bzw. den Abstand eines *read*-Paares übersteigt (Imelfort & Edwards 2009). Die im Pflanzenreich besonders unter vegetativ vermehrten und auskreuzenden Spezies weit verbreitete Heterozygotie verursacht in Bereichen, in denen die Haplotypen stark voneinander abweichen, einen Abbruch der Assemblierung (Michael & VanBuren 2015). Zudem ist es bei evolutionär jüngeren Duplikationsereignissen und großen Multigenfamilien praktisch unmöglich, zwischen paralogen Genen und zwei Allelvarianten desselben Gens zu unterscheiden (Hamilton & Robin Buell 2012). Noch komplizierter wird es bei polyploiden Pflanzenspezies, da die Redundanz in den Sequenzdaten, die durch das Vorhandensein verschiedener Subgenome zustande kommt, häufig zu einer Verschmelzung dieser bei der Assemblierung führt (Claros et al. 2012).

Lösungsansätze für diese Herausforderungen bei der Sequenzierung von Pflanzengenomen gibt es vielfach. Sie reichen von der bereits erwähnten Kombination verschiedener Sequenzierplattformen bis zu einer künstlichen Erniedrigung der Komplexität der Genome. Letzteres ist das Ziel verschiedener Verfahren, die unter dem Begriff *reduced representation sequencing* zusammengefasst werden können (Hamilton & Robin Buell 2012). Dazu zählt unter anderem ein Verdau mit methylierungssensitiven Restriktionsenzymen. Bei dieser Strategie macht man sich zu Nutze, dass repetitive Bereiche in eukaryotischen Genomen in der Regel hypermethyliert, Genbereiche hingegen hypomethyliert vorliegen (Emberton et al. 2005). Ein weiterer Ansatz nutzt die Renaturierungskinetik der DNA, um gezielt *single copy*-Bereiche des Genoms für die Sequenzierung zu isolieren (Yuan et al. 2003). Bei Pflanzenspezies, deren hoher Grad an Heterozygotie die Assemblierung erschweren würde, können Inzucht-Linien, die infolge wiederholter Selbstungen ein annähernd homozygotes Genom besitzen, als Grundlage für die Genomsequenzierung dienen (Jaillon et al. 2007). Bei polyploiden Arten kann eine der Sequenzierung vorausgehende physikalische Trennung der Chromosomen, beispielsweise mittels Durchflusszytometrie, helfen, die verschiedenen Subgenome getrennt voneinander zu sequenzieren (Paux et al. 2008). Nützlich kann in solchen Fällen auch eine vorangehende Sequenzierung nah verwandter diploider Vorgängerspezies sein, die aufgrund ihrer Syntänie Anhaltspunkte für die Assemblierung des polyploiden Genoms liefern können (Ling et al. 2013; Shulaev et al. 2011). Neben den reinen Sequenzdaten spielen Positionsinformationen von *paired end*- oder *mate pair*-Sequenzierungen eine immer größere Rolle, um die zahlreichen *contigs* zu einer größeren zusammenhängenden Sequenz, genannt *scaffold*, zusammenzufügen. Strategien zur Erstellung größerer Sequenzgerüste umfassen darüber hinaus oft die Verwendung genetischer Marker und eine optische *in situ*-Kartierung an den Chromosomen (Huang et al. 2009; Shearer et al. 2014).

Letztendlich sind jedoch weitere Fortschritte sowohl auf Seiten der Sequenziertechnologien als auch auf Seiten der Bioinformatik nötig, um der Komplexität der Pflanzengenome in ihrer gesamten Vielfalt Herr zu werden. Sequenzierplattformen der dritten Generation, wie die 2010 erstmalig erschienenen Geräte der Firmen Pacific Bioscience und Ion Torrent stellen dabei eine vielversprechende Entwicklung dar (Rusk 2011). Gerade ersteres eignet sich mit maximalen Leseweiten von über 20 kb sehr gut für die *de novo* Sequenzierung von Pflanzengenomen und die Überarbeitung bereits publizierter *draft genomes* (VanBuren et al. 2015; English et al. 2012).

1.5 Zielsetzung

Das Ziel der vorliegenden Arbeit ist eine umfassende Analyse des Genoms der Wilden Weinrebe. Mit den DNA-Sequenzen der Rebsorten Pinot Noir (Jaillon et al. 2007; Velasco et al. 2007), Tannat (Da Silva et al. 2013) und Sultanina (Di Genova et al. 2014) wurde das Rebengenom zwar schon mehrfach publiziert, jedoch beschränkten sich die Analysen ausschließlich auf die Subspezies der Edlen Weinrebe *Vitis vinifera* subsp. *vinifera*. Das Genom der Wilden Weinrebe *Vitis vinifera* subsp. *sylvestris* wurde hingegen bisher nicht näher charakterisiert.

Zur Entschlüsselung des Genoms sollen *Next-Generation Sequencing*-Technologien verwendet werden. Besonders im Fokus der Genomanalyse soll die genetische Diversität der Wilden Weinrebe stehen. Daher werden neben Einzelindividuen auch Pools von Wildrebenpopulationen sequenziert. Für das besondere Interesse an der genetischen Diversität der Wilden Weinrebe gibt es zwei Gründe. Erstens handelt es sich bei der Wilden Weinrebe um eine vom Aussterben bedrohte Wildpflanze. Ein Verschwinden der Wilden Weinrebe würde einen unwiederbringlichen Verlust der Biodiversität in der Gattung *Vitis* bedeuten. Daher soll im Rahmen dieser Arbeit die vorhandene Diversität in verschiedenen Wildrebenpopulationen ermittelt und Rückschlüsse auf Populationsstruktur und Evolutionsgeschichte gezogen werden. Zweitens stellt die Wilde Weinrebe als naher Verwandter der Kulturreben eine wichtige genetische Ressource für die Rebenzüchtung dar. Die Identifizierung Züchtungs-relevanter Kandidatengene in der Wilden Weinrebe soll daher ein weiterer Aspekt der vorliegenden Arbeit sein. Hierfür sollen molekulare Fußspuren der Selektion im Genom der Wilden Weinrebe ausfindig gemacht werden.

2 Material & Methode

2.1 Pflanzenmaterial

Als Ausgangsmaterial für die im Rahmen dieser Arbeit durchgeführten Experimente dienten junge Blätter der in Tabelle 6 dargestellten Individuen von *Vitis vinifera* subsp. *sylvestris* sowie *Vitis vinifera* subsp. *vinifera*. Die Entnahme der Blätter erfolgte bevorzugt im Frühjahr auf den Versuchsflächen des Julius Kühn-Instituts Geilweilerhof (Prof. Töpfer, Institut für Rebenzüchtung) sowie der Hochschule Geisenheim (Prof. Rühl, Institut für Rebenzüchtung) und im Botanischen Garten des Karlsruher Instituts für Technologie (Prof. Nick, Botanisches Institut). In Ausnahmefällen wurde jahreszeitenbedingt anstelle junger Blätter Rebholz geerntet. Das Pflanzenmaterial wurde auf Trockeneis nach Mainz transportiert und bis zur weiteren Verwendung bei -80 °C tiefgefroren.

Phänotypische Analysen der Blütenmorphologie erfolgten stets zu Beginn der Weinblüte in den ersten zwei Juniwochen an der Hochschule Geisenheim. Betrachtet wurden Blüten mit vollständig entfalteten Blütenorganen kurz nach Abwurf des Blütenkäppchens. Zur Dokumentation wurden die Blütenstände vom Rebstock entfernt und unmittelbar danach vor einem weißen Hintergrund fotografiert.

2.2 Puffer und Lösungen

Agarplatten	7,5 g Agar Agar
	500 ml LB-Medium
	5 ml Ampicillin-Stammlösung
	500 µl IPTG-Lösung
	500 µl X-Gal-Lösung
Ampicillin-Stammlösung	10 mg/ml Ampicillin
	in HPLC-Wasser
CI (24:1)	98 % Chloroform
	2 % Isoamylalkohol

Tabelle 6: Pflanzenmaterial

(Fortsetzung auf der folgenden Seite)

	Individuum	Entnahmeort	Herkunft	Falls vorhanden: Identifikationsnr.	
				VIVC	Accession name / IPEN
Einzelseq	Weißer Heunisch	Geisenheim	Geisenheim, Deutschland	n/a	Klon 8, M-2-18
	L-17-12-2	Geisenheim	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	13693	DEU454-L-17-12-2
	Hördt29	Geilweilerhof	Hördt, Deutschland	24049	DEU098-2013-001
	Kaukasus	KIT	Kaukasus	n/a	n/a
Pool Ketsch	Ketsch 2-29	Geilweilerhof	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	13683	DEU098-1980-103
	Ketsch 3	Geisenheim	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	13684	DEU454-L-20-12-2
	Ketsch 6	Geilweilerhof	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	13686	DEU098-2012-051
	Ketsch 7	Geisenheim	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	13690	DEU454-L-18-13-2
	Ketsch 10	Geisenheim	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	13691	DEU454-L-17-14-2
	Ketsch 16	KIT	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	n/a	DE-1-UNKAR-2011-6196
	Ketsch 21	Geilweilerhof	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	22638	n/a
	Ketsch 23	Geilweilerhof	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	13694	DEU098-2012-052
	Ketsch 27	Geilweilerhof	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	13695	DEU098-2012-053
	Ketsch 28	Geilweilerhof	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	13697	DEU098-2012-055
	Ketsch 32	Geisenheim	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	13698	DEU454-L-19-12-2
	Ketsch 34	Geilweilerhof	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	13699	DEU098-2013-051
	Ketsch 36	Geilweilerhof	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	22635	DEU098-2008-031
	Ketsch 60	Geilweilerhof	Naturschutzgebiet "Ketscher Rheininsel", Deutschland	22642	DEU098-2008-039
Pool Frankreich	Frankreich Abbadia H	Geilweilerhof	Zone naturelle, Pyrénées-Atlantiques, Frankreich	n/a	FRA139-8500Mtp257
	Frankreich Carranques 3	Geilweilerhof	Massif des Albères, en zone naturelle, Frankreich	25128	DEU098-2016-014
	Frankreich Bois Bourdet 1	Geilweilerhof	Cherves-Richemont, Frankreich	23590	DEU098-2012-089
	Frankreich PSL11	Geilweilerhof	Zone naturelle, Hérault, Pic Saint-Loup, Frankreich	23595	DEU098-2012-094
	Frankreich PSL12	Geilweilerhof	Zone naturelle, Hérault, Pic Saint-Loup, Frankreich	23596	DEU098-2012-095
	Frankreich PSL14	Geilweilerhof	Zone naturelle, Hérault, Pic Saint-Loup, Frankreich	25129	DEU098-2016-015
	Frankreich Grésigne 1	Geilweilerhof	Forêt de Grésigne Région de Gaillac Tarn, Frankreich	23592	DEU098-2012-091
	Frankreich l'Escale 1	Geilweilerhof	Zone naturelle, Ariège, Frankreich	23594	DEU098-2012-093
	Frankreich Grésigne 2	Geilweilerhof	Forêt de Grésigne Région de Gaillac Tarn, Frankreich	23593	DEU098-2012-092
	Frankreich PSL2	Geilweilerhof	Zone naturelle, Hérault, Pic Saint-Loup, Frankreich	23597	DEU098-2012-096

	Individuum	Entnahmeort	Herkunft	Falls vorhanden: Identifikationsnr.	
				VIVC	Accession name / IPEN
Pool Spanien	Spanien CA 2.5	Geilweilerhof	Río Majaceite (El Bosque) - Cádiz, Spanien	23573	DEU098-2012-072
	Spanien SE 1.4	Geilweilerhof	Los Melonares (Castilblanco de los arroyos) - Sevilla, Spanien	23587	DEU098-2012-086
	Spanien SE 2.7	Geilweilerhof	Ribera del Huéznar (Cazalla de la Sierra) - Sevilla, Spanien	23588	DEU098-2012-087
	Spanien O 1.2	Geilweilerhof	Río Cares (Vivoli) - Asturias, Spanien	23583	DEU098-2012-082
	Spanien S 1.4	Geilweilerhof	Cueva Covalanas (Ramales de la Victoria) - Cantabria, Spanien	23585	DEU098-2012-084
	Spanien NA 1.2	Geilweilerhof	Oronoz/Mugaire - Navarra, Spanien	23582	DEU098-2012-081
	Spanien SS 2.5	Geilweilerhof	Talaimendi (Zarautz) - Guipúzcoa, Spanien	n/a	ESP080-BGVCAM3223
	Spanien BU 1.4	Geilweilerhof	Peña Angulo, refugio (Artziniega) - Burgos, Spanien	23572	DEU098-2012-071
	Spanien CC 1.2	Geilweilerhof	N-630, km 444 (Casas del Monte) - Cáceres, Spanien	23574	DEU098-2012-073
	Spanien CR 1.1	Geilweilerhof	Río las Yeguas (Fuencaliente) - Ciudad Real, Spanien	23575	DEU098-2012-074
	Spanien H 7.8	Geilweilerhof	Doñana, Las Algaidas de Meloncillo y Acrizal (Almonte) - Huelva, Spanien	23577	DEU098-2012-076
	Spanien MA 2.7	Geilweilerhof	Río Turón 1 (El Burgo) - Málaga, Spanien	23579	DEU098-2012-078
	Spanien MA 2.8	Geilweilerhof	Río Turón 1 (El Burgo) - Málaga, Spanien	23580	DEU098-2012-079
	Spanien O 1.6	Geilweilerhof	Río Cares (Vivoli) - Asturias, Spanien	23584	DEU098-2012-083
	Spanien S 2.2	Geilweilerhof	Cueva Covalanas (Ramales de la Victoria) - Cantabria, Spanien	23586	DEU098-2012-085
Pool Pisa	Pisa 04	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7420
	Pisa 08	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7422
	Pisa 10	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7424
	Pisa 11	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7426
	Pisa 14	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7428
	Pisa 17	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7429
	Pisa 18	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7430
	Pisa 23	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7432
	Pisa 26	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7434
	Pisa 52	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7435
	Pisa 53	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7436
	Pisa 64	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7438
	Pisa 66	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7439
	Pisa 70	KIT	Università di Pisa, Pisa, Italien	n/a	IT-1-UNKAR-2012-7419

CTAB-Extraktionspuffer	2 % CTAB
	20 mM Na_2EDTA
	100 mM Tris
	1,4 M NaCl
	33,6 % Harnstoff
	in VE-Wasser
10x Dialysepuffer	3 M NaCl
	0,25 M Tris
	0,1 M Na_2EDTA
	in VE-Wasser, pH 7,6
10x E-Puffer	0,36 M Tris
	0,3 M NaH_2PO_4
	0,1 M Na_2EDTA
	in VE-Wasser
Ethidiumbromid-Färbelösung	0,1 % (v/v) EtBr-Stammlösung
	in 1x E-Puffer
Ethidiumbromid-Stammlösung	0,5 % (v/v) EtBr
	in 1x E-Puffer
IPTG-Lösung	23,8 mg IPTG
	in 500 µl DMF lösen
Ladepuffer	10 mM Tris-HCl
	0,15 % Orange G Farbstoff
	0,03 % Xylencyanol FF Farbstoff
	60 % Glycerin
	60 mM EDTA
LB-Medium	10 g Trypton
	5 g Hefeextrakt
	5 g NaCl
	mit VE-Wasser auf 500 ml auffüllen

Qubit™ Arbeitslösung	0,5 % (v/v) Qubit™ dsDNA HS Reagenz in Qubit™ dsDNA HS Puffer
10x TBE-Puffer	108,0 g Tris 55,0 g Borsäure 8,3 g Na_2EDTA mit VE-Wasser auf 1 l auffüllen
X-Gal-Lösung	50 mg X-Gal in 500 µl DMF lösen

2.3 Molekularbiologische Methoden

2.3.1 Isolierung von Nukleinsäuren

2.3.1.1 DNA-Extraktion aus jungen Blättern

Die Isolierung der DNA aus jungen Blättern erfolgte mit dem NucleoSpin® Plant II Kit (Macherey-Nagel, Düren, Deutschland) unter Verwendung des Lysepuffers PL1. Die Angaben des Herstellers wurden folgendermaßen modifiziert: Die Lyse bei 65 °C wurde auf insgesamt 30 min ausgedehnt, nach jeweils 10 und 20 min wurde die Inkubation durch kurzes Invertieren der Proben unterbrochen. Im Anschluss wurde das Lysat für 2 min bei 11.000 g und RT zentrifugiert und nur der Überstand für die weiteren Extraktionsschritte verwendet. Die isolierte DNA wurde in 50 µl vorgewärmtem HPLC-H_2O eluiert.

2.3.1.2 DNA-Extraktion aus Rebholz

0,5 g Rebholz wurden mit dem Skalpell entrindet und in 0,5 cm große Stücke zerkleinert. Diese wurden gemeinsam mit 5 ml CTAB-Extraktionspuffer und einer Spatelspitze PVPP in eine Extraktionstüte (Extraction bags «universal» von Bioreba, Reinach, Schweiz) gegeben und mit einem Hammer homogenisiert. Das Gemisch wurde in Eppendorfgefäße überführt, mit 1 µl RNase A (Peqlab, Erlangen, Deutschland) versetzt und für 30 min bei 65 °C in einem Thermomixer (Eppendorf, Hamburg, Deutschland) inkubiert. Anschließend erfolgte eine zehnminütige Zentrifugation bei 14.000 g und RT. Zum Entfernen von Proteinen wurden mit dem Überstand drei CI-Extraktionen durchgeführt. Hierzu wurde der Überstand mit 1 Vol. CI versetzt und nach kurzem Invertieren für 10 min bei 14.000 g und RT zentrifugiert. Die

obere Nukleinsäure-haltige Phase wurde in ein neues Eppendorfgefäß überführt. Abschließend wurde die DNA gefällt (vgl. Kapitel 2.3.2) und in 50 µl HPLC-H_2O gelöst.

2.3.2 Fällung von Nukleinsäuren

Die Proben wurden zwecks Fällung mit 1/10 Vol. 10x Dialysepuffer und 2,5 Vol. EtOH abs. versetzt und invertiert. Die DNA wurde für eine Stunde bei -20 °C gefällt und mittels Zentrifugation für 30 min bei 4 °C und 14.000 g pelletiert. Der Überstand wurde verworfen und das Pellet mit 500 µl 70 % EtOH gewaschen, um Salzrückstände zu entfernen. Nach zehnminütiger Zentrifugation bei RT und 14.000 g wurde der Überstand erneut verworfen, das DNA-Präzipitat getrocknet und je nach Folgeanwendung in 10–50 µl HPLC-H_2O gelöst.

2.3.3 Konzentrationsbestimmung von Nukleinsäuren

Die Konzentration von Nukleinsäuren wurde entweder photometrisch anhand 1 µl Lösung mit dem NanoDrop® ND-1000 (Peqlab, Erlangen, Deutschland) bestimmt oder fluorometrisch mit dem Qubit™ (Invitrogen, Carlsbad, Kalifornien, USA) und dem Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, Kalifornien, USA) ermittelt. Bei Messungen mit dem Qubit™ Fluorometer fand das Qubit™ dsDNA HS Assay Kit Verwendung. Die Proben wurden hierbei 1:200 oder 1:100 mit der frisch hergestellten Qubit™ Arbeitslösung verdünnt und die Messung nach Herstellerangaben durchgeführt. Der Agilent 2100 Bioanalyzer kombiniert die Konzentrationsbestimmung mit einer gelelektrophoretischen Bestimmung der Größe. Daher eignet er sich besonders zur Qualitätsbewertung von Bibliotheken für die Hochdurchsatzsequenzierung. Hierfür wurde das Agilent High Sensitivity DNA Kit nach Herstellerangaben verwendet.

2.3.4 Gelelektrophoretische Auftrennung von DNA

Zur Auftrennung von Nukleinsäuren gemäß ihres Molekulargewichts wurde die Agarose-Gelelektrophorese nach Sambrook und Russell (2006) eingesetzt. Hierbei wurde die GENterphorese™-Apparatur der Firma Genterprise (Mainz, Deutschland) mit 1x TBE als Laufpuffer verwendet. Der Agaroseanteil der Gele betrug je nach zu trennender Fragmentgröße zwischen 0,8 % und 2 %. Für die Größenbestimmung wurden die GeneRuler™ 100 bp Plus DNA Leiter und der λ DNA/Hind III Marker (beide Thermo Scientific, Waltham, Massachusetts, USA) als Molekulargewichtsstandard eingesetzt. 6x Ladepuffer beschwerte die DNA-Proben und ermöglichte eine Sichtkontrolle der

Laufweite während der Gelelektrophorese. Die Elektrophorese erfolgte bei RT und einer Stromstärke von 130 mA für ca. 20–30 min. Im Anschluss wurde das Gel für 2 min in einer EtBr-Färbelösung inkubiert, kurz gewässert und auf einem UV-Transilluminator bei einer Wellenlänge von 312 nm mithilfe des Carestream Gel Logic 112 Imaging Systems (Carestream Health, Rochester, New York, USA) dokumentiert.

2.3.5 Polymerasekettenreaktion

Die Polymerasekettenreaktion dient der Amplifikation eines spezifischen von zwei Primern flankierten DNA-Abschnitts mithilfe einer thermostabilen DNA-Polymerase (Mullis et al. 1986). Die im Rahmen dieser Arbeit verwendeten Primer (Sigma-Aldrich, St. Louis, Missouri, USA) sind in Tabelle 7 aufgelistet. Sie besaßen optimalerweise eine Länge von 18–25 bp und eine Schmelztemperatur von 57–62 °C. Die Schmelztemperatur und die Tendenz, unerwünschte Homo- bzw. Heterodimere auszubilden, wurden mit den Online-Tools IDT OligoAnalyzer 3.1 (IDT, Coralville, Iowa, USA) und dem Oligonucleotide Properties Calculator (Kibbe 2007) berechnet. Vor Verwendung wurden die Primer mit HPLC-H_2O auf eine Konzentration von 10 pmol/µl eingestellt.

Tabelle 7: Übersicht der verwendeten Primer

	Name	Sequenz	T_m [°C]	Quelle
Farblocus	d3	CCTGCAGCTTTTTCGGCATCT	61,2	Lijavetzky et al. 2006
	e1	GTCTTTCGCTTGCCAACTGT	58,4	Lijavetzky et al. 2006
	LTR5f	AGAAGGGGATCCTCCTGGTA	60,5	This et al. 2007
	MybA1_spez.	GCATCTCTCCAGAAGCCG	58,4	Sabine Fischer
	MybA2_spez.	CACAGAAAAAGGGAACACATTC	58,4	Sabine Fischer
	Pf	GTCCCAAGCAACAGATGGAT	58,4	This et al. 2007
Sex	VSVV010_for	AGTGCTCACTTTTCCTTGTGAA	58,4	Picq et al. 2014
	VSVV010_rev	CATGAATCAGCAGTGCATTT	54,3	Picq et al. 2014
Transp.	IRAP_Tvv1_Fa	TCCAGCTTGAGGGGGAGTGT	62,5	D'Onofrio et al. 2010
	IRAP_Gret1_Fc	CCATGGCTAACAAAACCATC	56,4	Castro et al. 2012
	PBS_F0100	TAGGTCGGAACAGGCTCTGATACCA	67,4	Kalendar et al. 2008
Allelfreq.	Ke_01_101_313644_for	CTTTTCTGTCTCTGCGATTAC	57,5	Sabine Fischer
	Ke_01_101_313644_rev	GATGATCCTTAGACTAGTTCC	57,5	Sabine Fischer
	Sp_01_12_336274_for	ATGAGGACAAATGCCAGATAC	57,5	Sabine Fischer
	Sp_01_12_336274_rev	GCCTTGGTTGATAGGTGAC	57,5	Sabine Fischer

Die erste Spalte gibt an, in welchem Projekt die jeweiligen Primer verwendet wurden. Farblocus: Analyse des genetischen Locus, der für die Beerenfarbe verantwortlich ist; Sex: Bestimmung des Geschlechts der Weinreben; Transp.: Transposon-basierte molekulare Marker; Allelfreq.: Überprüfung der Allelfrequenzen in der Pool-Sequenzierung.

Als thermostabile Polymerase diente die GoTaq® DNA-Polymerase (5 U/µl) mit dem zugehörigen 5x Reaktionspuffer (beides Promega, Madison, Wisconsin, USA). Ein beispielhafter PCR-Ansatz ist in Tabelle 8 dargestellt. Die Amplifikation erfolgte in einem PTC DNA Engine® Thermal Cycler (Bio-Rad, Hercules, Kalifornien, USA) unter den in Tabelle 9 dargestellten Reaktionsbedingungen. In Ausnahmefällen wurde auf Touchdown- (Don et al. 1991) und long-range-PCR-Protokolle (Davies & Gray 2002) zurückgegriffen.

Tabelle 8: PCR-Ansatz

Template DNA (25 ng/µl)	1 µl
GoTaq® DNA Polymerase (5 U/µl)	0,3 µl
dNTPs (10 mM each)	1 µl
Primer for (0,01 mM)	2 µl
Primer rev (0,01 mM)	2 µl
$MgCl_2$ (25 mM)	4 µl
5x GoTaq® Flexi Puffer	10 µl
HPLC-H_2O	29,7 µl
Gesamtvolumen	50 µl

Tabelle 9: PCR-Programm

Initiale Denaturierung	94 °C	4:00 min	
Denaturierung	94 °C	0:30 min	
Annealing	56–65 °C	0:30 min	40 x
Elongation	72 °C	1 min/kb	
Finale Elongation	72 °C	10 min	
	4 °C	∞	

2.3.6 Aufreinigung von PCR-Produkten

Falls das erfolgreich amplifizierte PCR-Fragment direkt zur Kapillarsequenzierung eingesetzt werden sollte, erfolgte zunächst ein halbstündiger Verdau mit 10 U Exonuclease I und 0,9 U Fast AP (beides Thermo Scientific, Waltham, Massachusetts, USA) bei 37 °C. Die Enzyme wurden im Anschluss durch eine fünfzehnminütige Inkubation bei 72 °C inaktiviert. Entstanden neben dem beabsichtigten Produkt bei der PCR unspezifische Nebenprodukte, wurde die gewünschte Bande mittels Skalpell aus dem Agarosegel ausgeschnitten. Anschließend wurde die DNA aus dem Gelstück mithilfe des GelElute™ Gel Extraction Kit (Sigma Aldrich, St. Louis, Missouri, USA) nach Angaben des Herstellers wiedergewonnen.

2.3.7 Klonierung von PCR-Produkten

Für die Klonierung von PCR-Produkten wurde das in Abbildung 5 gezeigte pGEM®- T easy-Vektorsystem (Promega, Madison, Wisconsin, USA) gemäß Herstellerangaben verwendet.

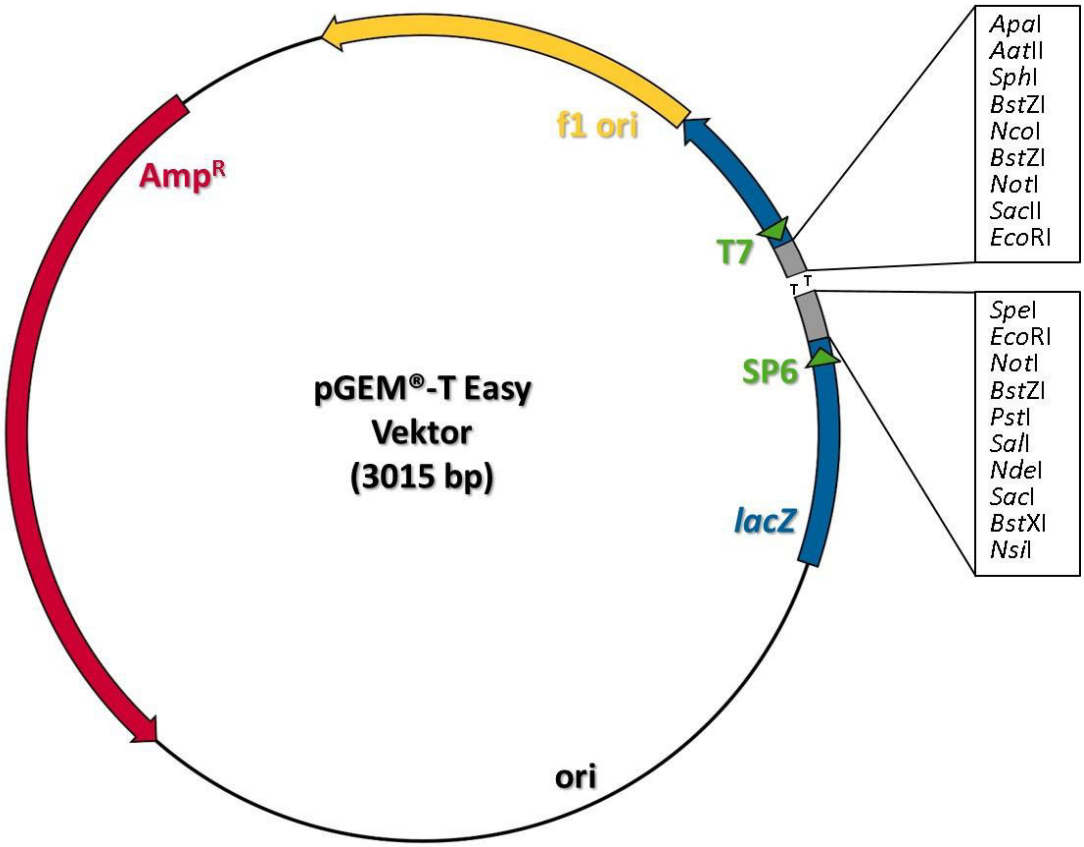

Abbildung 5: Vektorkarte von pGEM®-T Easy
Dargestellt ist das für Klonierungsexperimente verwendete pGEM®-T easy Vektorsystem mit einem *high-copy*-Replikationsursprung und dem β-Lactamasegen, das seinem Träger eine Ampicillinresistenz verleiht (Amp^R). Für eine Blau-Weiß-Selektion befindet sich die multiple Klonierungsstelle (MCS, *multiple cloning site*) innerhalb des lacZ'-Gens. Der linearisierte Vektor besitzt an den 3'-Enden jeweils ein überhängendes Thymidin, was eine T/A-Klonierung ermöglicht. Grüne Dreiecke markieren die Primerbindestellen der für die Sanger-Sequenzierung verwendeten Primer T7 und SP6.

Die Ligation erfolgte in einem 10 µl Ansatz (Tabelle 10) über Nacht bei 4 °C. Anschließend wurde der Ligationsansatz gefällt (vgl. Kapitel 2.3.2) und in 9 µl HPLC-H_2O gelöst.

Tabelle 10: Ligationsansatz

PCR-Produkt	3–7 µl
pGEM®- T easy-Vektor	1 µl
T4 DNA-Ligase	1 µl
10x DNA-Ligase-Puffer	1 µl
HPLC-H_2O	0–4 µl
Gesamtvolumen	10 µl

Für die Transformation wurde der *E. coli* Stamm DH10B mit dem Genotyp F− *araD139* Δ(*ara, leu*)7697 Δ*lacX74 galU galK rpsL deoR* Φ80d*lacZ*ΔM15 *endA1 nupG recA1 mcrA* Δ(*mrr*

hsdRMS mcrBC) verwendet (Grant et al. 1990). Die Bakterienzellen wurden in der log-Phase ihres Wachstums durch wiederholte Waschschritte mit sterilem kalten HPLC-H_2O gefolgt von sterilem kalten 10 %-igem Glycerin elektrokompetent gemacht und bei -80 °C gelagert. Die Transformation der elektrokompetenten Zellen mit 1 µl des gefällten Ligationsansatzes erfolgte durch Elektroporation mit einem MikroPulser™ (Bio-Rad, Hercules, Kalifornien, USA). Die transformierten Zellen wurden in 1 ml vorgewärmten L-Medium resuspendiert und zwecks Expression der Ampicillinresistenz für eine Stunde unter Schütteln bei 37 °C inkubiert. Im Anschluss wurden 100–500 µl des Transformationsansatzes auf X-Gal, IPTG und Ampicillin enthaltende Agarplatten ausplattiert und über Nacht bei 37 °C bebrütet. Bewachsene Agarplatten wurden bis zu vier Wochen bei 4 °C gelagert.

2.3.8 Plasmid-Präparation

Für die Plasmid-Präparation wurden über Nacht-Kulturen von weißen potentiell rekombinanten Kolonien in 5 ml L-Medium mit 1 % Ampicillin angelegt. Die Anzucht der Klone erfolgte unter Schütteln bei 37 °C über Nacht. Die Plasmide wurden mit dem GenElute™ HP Plasmid Miniprep Kit (Sigma Aldrich, St. Louis, Missouri, USA) nach Angaben des Herstellers isoliert. Die Integratgröße wurde anhand eines Restriktionsverdaus (Tabelle 11) mit EcoRI (10 U/µl, Thermo Scientific, Waltham, Massachusetts, USA) und anschließender Agarose-Gelelektrophorese (vgl. Kapitel 2.3.4) überprüft.

Tabelle 11: Restriktionsverdau

Plasmid	2 µl
EcoRI	1 µl
10x Restriktionspuffer	2 µl
HPLC-H_2O	15 µl
Gesamtvolumen	20 µl

2.3.9 Sanger-Sequenzierung

Die Sanger-Sequenzierungen von aufgereinigten PCR-Produkten (vgl. Kapitel 2.3.6) und klonierten Fragmenten (vgl. Kapitel 2.3.7) wurden von der Firma Genterprise (Mainz, Deutschland) durchgeführt. Hierfür wurden 50–200 ng der aufgereinigten PCR-Produkte mit einem zugehörigen Primer bzw. 400–700 ng Plasmid-DNA mit T7- oder SP6-Primer eingesetzt. Die im .ab1-Format erhaltenen Sequenzdaten wurden mit dem Programm FinchTV (Geospiza, Seattle, Washington, USA) visuell kontrolliert. Abschnitte schlechter Qualität in Anfangs- und Endbereich der Sequenz wurden entfernt. Sequenzen von

Plasmidsequenzierungen wurden mit dem Online-Tool VecScreen (NCBI, Bethesda, Maryland, USA) von Vektor-Kontamination bereinigt. Die weitere bioinformatische Auswertung der editierten Sequenzen erfolgte mit der CLC Main Workbench 6.9.1 (CLC bio, Aarhus, Dänemark). Je nach Fragestellung wurden hierbei Assemblierungen, Alignments mit bekannten Referenzsequenzen, Dot-Plot-Analysen oder BLAST-Suchen in verschiedenen Datenbanken (z.B. *Nucleotide collection* nr/nt von NCBI) durchgeführt.

2.3.10 Genotypisierung von Weinreben

Die Genotypisierung wurde im Julius Kühn-Institut Geilweilerhof (Dr. Erika Maul, Institut für Rebenzüchtung) durchgeführt. Auf Basis der vom europäischen Konsortium *GrapeGen06* erarbeiteten SSR-Marker (*simple sequence repeats*) wurden 9 Loci ausgewählt, die für die Unterscheidung von *Vitis vinifera* subsp. *vinifera* und *Vitis vinifera* subsp. *sylvestris* als informativ gelten (Ledesma-Krist et al. 2013). Bei den untersuchten Loci handelt es sich um VVS2, VVMD5, VVMD7, VVMD25, VVMD27/ZAG47, VVMD28, VVMD32, VrZAG62 und VrZAG79. Die Durchführung erfolgte gemäß This et al. (2004).

2.4 *Next-Generation Sequencing*

Für die Hochdurchsatzsequenzierung wurden *paired-end*-Bibliotheken aus genomischer DNA von Wilden Weinreben und einer Kulturreben hergestellt. Für die Sequenzierung von Pools wurde vorab die DNA von bis zu 15 Individuen (Tabelle 6) mittels Qubit™-Messung (vgl. Kapitel 2.3.3) quantifiziert und in gleichen Mengenverhältnissen vereint. Die Fragmentierung der hochmolekularen genomischen DNA erfolgte mit der *Adaptive Focused Acoustics*™ Technologie von Covaris® (Woburn, Massachusetts, USA). Die Bibliotheken wurden mit dem TruSeq DNA LT Sample Prep Kit (Illumina, San Diego, Kalifornien, USA) nach Herstellerangaben erstellt. Eine Ausnahme stellten die Proben „Weißer Heunisch“ und „L-17-12-2“ dar, hier erfolgte die Erstellung der Bibliotheken wie von Fischer (2012) beschrieben. Die Illumina-Sequenzierung wurde als 100 bp-*paired-end*-Lauf auf dem HiSeq2000/2500 (Illumina, San Diego, Kalifornien, USA) am Nukleinsäure-Analyse-Zentrum (IMSB, Mainz, Deutschland) durchgeführt.

2.5 Bioinformatische Methoden

2.5.1 Perl-Skripte

Bei der Analyse der NGS-Daten fanden zahlreiche Perl-Skripte Anwendung. Sie sind in Tabelle 12 mit einer kurzen Beschreibung ihrer Funktion aufgelistet. Alle im Rahmen dieser Arbeit verwendeten Perl-Skripte wurden von Benjamin Rieger (IMSB, Mainz, Deutschland) verfasst. Sie sind im elektronischen Anhang zu finden. Als Perl-Interpreter diente ActivePerl 5.16.3 (ActiveState, Vancouver, Kanada).

Tabelle 12: Übersicht aller verwendeten Perl-Skripte

Perl-Skript	Beschreibung
qseq_pe_tag_sort_list.pl	sortiert Sequenzen einer .qseq-Datei nach Adapterindizes
qseq2fastq.pl	konvertiert Sequenzdateien aus dem .qseq-Format in das .fastq-Format
fastq_integrity.pl	überprüft die Integrität der erzeugten .fastq-Sequenzdateien
seq_info.pl	analysiert Sequenzdateien hinsichtlich Anzahl der Sequenzen und Basen, durchschnittliche Länge der Sequenzen, N50, N90, Länge der kürzesten und längsten Sequenz, absolute und relative Basenzusammensetzung, Zahl der Sequenzen mit einer durchschnittlichen Qualität ≥ 20/30/40
fastq_average_qual.pl	berechnet die durchschnittliche Qualität aller Sequenzen eines Datensatzes
fastq_paired_filter.pl	filtert und bereinigt Rohdaten im .fastq-Format
fasta_n_cut_with_n_length.pl	zerlegt eine .fasta-Datei an Positionen mit einer frei wählbaren Anzahl undefinierter Nukleotide (N) in getrennte *contigs*
sliding_window.pl	berechnet auf Basis einer Polymorphismen-Tabelle die gepoolte Heterozygosität (H_P) in das Genom entlanggleitenden Fenstern
sliding_window_parse1.pl	fasst die H_P-Werte eines Chromosoms in einer Datei zusammen
sliding_window_parse2.pl	fasst die H_P-Werte eines Genoms in einer Datei zusammen
ZH_P_calculator.pl	führt eine Z-Transformation durch, um auf Basis der H_P-Werte ZH_P-Werte zu berechnen

2.5.2 Programme, Online-Tools und Datenbanken

Neben den angesprochenen Perl-Skripten wurden weitere bioinformatische Hilfsmittel zur Datenauswertung verwendet. Tabelle 13 fasst diese Programme, Online-Tools und Datenbanken samt Herkunft und Funktion zusammen.

Tabelle 13: Liste verwendeter Programme, Online-Tools und Datenbanken

Name	Quelle und Beschreibung
BLAST	Altschul et al. (1990) https://blast.ncbi.nlm.nih.gov/Blast.cgi Das *Basic Local Alignment Search Tool* vergleicht biologische (d. h. Aminosäure- oder Nukleotid-) Sequenzen mit Sequenzen aus Datenbanken in Form von lokalen Alignments
CLC Genomics Workbench	Qiagen, Hilden, Deutschland; ehemals CLC bio kommerzielle Software zur Analyse von NGS-Daten, verwendet für den Import der Illumina-Daten, deren Kartierung gegen das Referenzgenom, die Identifizierung der Polymorphismen sowie die Annotation der Kandidatenregionen
FastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ Programm zur Qualitätskontrolle von Sequenzdaten verschiedener Hochdurchsatzsequenzierungsplattformen
FinchTV	Perkin Elmer, Waltham, Massachusetts, USA; ehemals Geospiza Programm zum Betrachten und Editieren von Sequenzier-Chromatogrammen
OligoAnalyzer 3.1	https://eu.idtdna.com/calc/analyzer Online-Tool zum Berechnen von Primereigenschaften
Oligonucleotide Properties Calculator	Kibbe (2007) http://biotools.nubic.northwestern.edu/OligoCalc.html Online-Tool zum Berechnen der Schmelztemperatur von Primern
Pearson Korrelation	http://www.socscistatistics.com/tests/pearson/ Online-Tool zur Berechnung des Pearsonschen Korrelationskoeffizienten r
PLAZA 3.0	Proost et al. (2015) http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/ Online-Tool für komparative Genomanalysen in Pflanzen
Repbase Update	Bao et al. (2015) http://www.girinst.org/repbase/update/index.html Datenbank mit repetitiven Elementen eukaryotischer Genome
SRA	https://www.ncbi.nlm.nih.gov/sra/ Das *Sequence Read Archive* ist eine öffentliche Datenbank mit frei verfügbaren Sequenzdaten von *Next-Generation Sequencing* Projekten
TAIR	https://www.arabidopsis.org/ Datenbank mit genetischen und molekularbiologischen Informationen zum Modellorganismus *Arabidopsis thaliana*
VecScreen	https://www.ncbi.nlm.nih.gov/tools/vecscreen/ Online-Tool zur Identifizierung von Vektorkontamination in einer Sequenz
Venny 2.1	http://bioinfogp.cnb.csic.es/tools/venny/ Online-Tool zum Abgleich von bis zu vier Listen und zum Zeichnen von Venn-Diagrammen auf Basis dieser Listen

2.5.3 Verarbeitung der Rohdaten

Vor 2013 generierte Sequenzen lagen als Dateien im .qseq-Format vor. Innerhalb einer *Lane* der *Flowcell* konnten verschiedene Bibliotheken sequenziert werden, die sich durch die Indizes der ligierten Adapter unterschieden. Um die Sequenzen der einzelnen Bibliotheken zu trennen, wurden die Rohdaten mit dem Skript qseq_pe_tag_sort_list.pl nach den Adapterindizes sortiert. In einem folgenden Schritt wurden die .qseq-Dateien mit dem

Skript qseq2fastq.pl in das .fastq-Format umgewandelt und die Integrität der erzeugten .fastq-Dateien mit fastq_integrity.pl überprüft. Nach 2013 erstellte Sequenzen lagen dank des Programms bcl2fastq 1.8.4 (Illumina, San Diego, Kalifornien, USA) direkt als demultiplexte Dateien im .fastq-Format vor.

Die Rohdaten im .fastq-Format wurden mit dem Skript fastq_paired_filter.pl gefiltert. Die gewählten Parameter sind in Tabelle 14 zusammengefasst.

Tabelle 14: Filterparameter der NGS-Daten

Option	Parameter	Wert
seed	Mindestlänge der Adapter-Subsequenz, die eine Adapter-Filterung auslöst	9
mm	Maximalzahl an *Mismatches*, die eine Adapter-Subsequenz beim Alignment mit der zu filternden Sequenz aufweisen darf	0.1, 0.5
add-nuc	Anzahl der Nukleotide, die über den Adapter hinaus entfernt werden	1
ft5	Anzahl der Nukleotide, die am 5'-Ende entfernt werden	1
ft3	Anzahl der Nukleotide, die am 3'-Ende entfernt werden	5
ws	Fenstergröße des Qualitätsfilter-Algorithmus	12
qc	Qualitätswert, den die Positionen innerhalb eines Fensters im Durchschnitt mindestens erreichen müssen, um nicht entfernt zu werden	30
qo	ASCII offset	33
mn	maximal erlaubte Anzahl undefinierter Nukleotide (N) nach dem Filtern	0
no-tag	An-/Aus-Schalter der Adapter-Filterung	an
no-ft	An-/Aus-Schalter der 5'- und 3'-Filterung	an
no-tn	An-/Aus-Schalter der N-Filterung	an
no-qual	An-/Aus-Schalter der Qualitäts-Filterung	an
no-te	An-/Aus-Schalter des Filters, der Sequenzabschnitte mit schlechter Qualität in den Randbereichen eines *reads* entfernt	an
min	Mindestlänge, die eine Sequenz nach dem Filtern haben muss	30

Vor und nach dem Filterprozess wurde die Zusammensetzung und Qualität der Sequenzen mit den Perl-Skripten seq_info.pl und fastq_average_qual.pl sowie dem Programm FastQC analysiert.

2.5.4 *De novo* Assemblierung

Die *de novo* Assemblierung der gefilterten Daten wurde mit der CLC Assembly Cell 4.20 (CLC bio, Aarhus, Dänemark) durchgeführt. Der implementierte Algorithmus nutzt sogenannte de Bruijn Graphen für die Assemblierung (Miller et al. 2010). Die Länge der Subsequenzen („Wörter" oder „*kmers*"), in die die Sequenzen zu Beginn der Assemblierung zerlegt wurden, betrug 24 bp. Als „*bubble size*" wurden 200 bp gewählt. Die resultierenden *contigs* mussten eine Mindestlänge von 200 bp besitzen.

2.5.5 Analyse transposabler Elemente

Die transposablen Elemente von *Vitis vinifera*, die in den folgenden Analysen als Referenz dienten, wurden aus der Datenbank Repbase Update als fasta-Datei heruntergeladen (346 Einträge, Stand: 22.07.2014). Ergänzt wurde die Datenbank um Sequenzen von transposablen Elementen aus den Publikationen von Benjak et al. (2008) und Wenke et al. (2011), die nicht bei Repbase Update hinterlegt waren. Insgesamt standen 379 transposable Elemente der Weinrebe zur Verfügung. Die gefilterten Illumina-Daten wurden mit der „*Map reads to Reference*"-Funktion der CLC Genomics Workbench 5.5.1 mit Server Plugin (CLC bio, Aarhus, Dänemark) gegen die erstellte Transposon-Datenbank kartiert. Mindestens 90 % einer Sequenz mussten mit der Referenz zu 90 % übereinstimmen, damit sie an der jeweiligen Position kartiert wurde. Die weiteren Parameter wurden wie in Kapitel 2.5.5 beschrieben gewählt. Im Anschluss an die Kartierung wurden die Anzahl der kartierten *reads* sowie die Länge der Konsensus-Sequenz für jeden Eintrag extrahiert.

Für jedes transposable Element wurde ein RPKM-Wert gemäß untenstehender Formel berechnet. Diese Berechnung entstand in Anlehnung an die Quantifizierung der Genexpression in RNA-Seq-Experimenten (Mortazavi et al. 2008). Im vorliegenden Fall dient sie jedoch zur Normalisierung der Anzahl der kartierten *reads* auf die Länge des transposablen Elements (anstelle der Transkriptlänge) und der Datensatzgröße.

$$RPKM = \frac{\frac{Anzahl\ der\ kartierten\ reads}{\frac{Konsensuslänge}{1000}}}{\frac{Datensatzgröße}{1.000.000}}$$

Zur Umrechnung des RPKM-Werts in eine Kopienzahl musste bekannt sein, welcher RPKM-Wert mit einer Kopie eines genetischen Elements gleichzusetzen ist. Zu diesem Zweck wurden für jeden Datensatz die RPKM-Werte von drei *single copy*-Genen der Weinrebe (GCR1, ERG28, PMS1; Duarte et al. 2010) berechnet. Die Kartierung erfolgte in diesem Fall mit einer Stringenz von 0,95/0,95.

Die Kopienzahl der transposablen Elemente wurde gemäß untenstehender Formel berechnet, wobei X für den Datensatz-spezifischen durchschnittlichen RPKM-Wert der *single copy*-Gene steht.

$$Kopienzahl_{transposables\ Element} = \frac{RPKM_{transposables\ Element}}{X}$$

2.5.6 Kartierung gegen das Referenzgenom

Als Referenzsequenz für Kartierungen diente die Version 12X.2 des von Jaillon et al. (2007) publizierten Genoms des Pinot Noir Inzuchtklons PN40024. Die Sequenzen der 19 Chromosomen sowie des artifiziellen R-Konstrukts wurden als multifasta-Datei von der Homepage der „*Unité de Recherches en Génomique Info*“ (URGI)[1], einer Forschungseinheit des INRA, heruntergeladen. Die Kartierung der Illumina-Daten gegen das Referenzgenom wurde mit der „*Map reads to Reference*“-Funktion der CLC Genomics Workbench 5.5.1 mit Server Plugin (CLC bio, Aarhus, Dänemark) vorgenommen. Es wurde ein lokales Alignment durchgeführt. Diese Form des Alignments erlaubt das Vorhandensein einiger nicht kartierter Nukleotide am Anfang und Ende einer Sequenz. Für Fehlpaarungen zwischen der kartierten Sequenz und der Referenz wurden zwei Strafpunkte vergeben. Für eine Insertion oder Deletion gab es drei Strafpunkte. Mindestens 95 % einer Sequenz mussten mit der Referenz zu 95 % übereinstimmen, damit sie an der jeweiligen Position kartiert wurde. Falls Sequenzen an mehreren Positionen kartiert werden konnten, wurden sie auf diese Positionen zufällig verteilt und als unspezifisch gekennzeichnet. Der Abstand der *read*-Partner wurde vom Programm automatisch und für jeden Datensatz spezifisch kalkuliert. Kartierten die beiden Partner innerhalb des ermittelten Intervalls, wurden sie als *reads in pairs* klassifiziert. War der Abstand der beiden Partner größer bzw. kleiner oder befanden sie sich an weit entfernten Positionen wurden sie als *broken paired reads* geführt. Sequenzen, die nicht im Referenzgenom kartierten, wurden verworfen.

Im Fall der Pool-Sequenzierung wurde die Konsensussequenz der Kartierung als .fasta-Datei exportiert, wobei Positionen ohne Abdeckung mit undefinierten Nukleotiden (N) aufgefüllt wurden. Im Anschluss wurde die extrahierte Konsensussequenz mithilfe des Skripts fasta_n_cut_with_n_length.pl an Positionen mit Lücken größer 500 N in getrennte *contigs* zerlegt. Die resultierenden *contigs* dienten als Referenz für eine zweite Kartierung. Auf diese Art und Weise wurde sichergestellt, dass Bereiche stromauf- und abwärts größerer Lücken in den folgenden Analysen als getrennte genetische Abschnitte betrachtet und nicht fälschlicherweise zu einer Einheit zusammengefasst wurden.

[1] https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences

2.5.7 Identifizierung von Polymorphismen

Polymorphismen wurden mit der *„Quality-based Variant Detection*"-Funktion der CLC Genomics Workbench 5.5.2 (CLC bio, Aarhus, Dänemark) identifiziert. Sie basiert auf dem *„Neighborhood Quality Standard"* (NQS)-Algorithmus von Altshuler et al. (2000). Um als Polymorphismus gewertet zu werden, musste die zugrundeliegende Position eine Abdeckung von mindestens 10 jedoch maximal 250 Sequenzen aufweisen. Sequenzen, die bei der Kartierung als unspezifisch gekennzeichnet wurden, wurden für die Identifizierung von Polymorphismen ausgeschlossen, *broken paired reads* waren hingegen zulässig. Die betroffene Base musste einen Qualitätswert von mindestens 20, die flankierenden sieben Basen einen Wert von wenigstens 15 erreichen. In dieser sieben Basen umfassenden Umgebung durften nur zwei Lücken oder Fehlpaarungen auftreten. In Berücksichtigung der Tatsache, dass es sich bei den Sequenzdaten um das Resultat einer Pool-Sequenzierung handelte, wurde die maximale Zahl der zu erwartenden Allele auf 10 erhöht, sowie die minimale Allelfrequenz auf 5 % erniedrigt. Diese Minimalfrequenz musste nicht erreicht werden, falls mehr als fünf Sequenzen die zu identifizierende Variante unterstützen. Die Ergebnisse wurden in Form einer Variantentabelle (*variant table*) gespeichert und als Tab-separierte Textdatei exportiert.

2.5.8 Berechnung der gepoolten Heterozygosität (H_P)

Um molekulare Fußspuren der Selektion im Genom der Wilden Weinrebe zu identifizieren, wurde die gepoolte Heterozygosität (H_P) nach Rubin et al. (2010) gemäß untenstehender Formel berechnet.

$F\text{ür } ein\ Fenster\ mit\ l\ Loci\ gilt$:

$$H_P = \frac{2 \sum_{i=l}^{l} n_i \ \sum_{i=l}^{l} (N_i - n_i)}{(\sum_{i=l}^{l} n_i + \ \sum_{i=l}^{l} (N_i - n_i))^2}$$

$wobei$:

$N_i = Zahl\ der\ Reads\ am\ Locus\ i$

$n_i = Zahl\ der\ Reads\ des\ h\ddot{a}ufigsten\ Allels\ am\ Locus\ i$

Die Kalkulation der H_P-Werte erfolgte hierbei automatisiert durch das Skript sliding_window.pl auf Basis der Tab-separierten Textdatei der Variantentabelle (vgl. Kapitel 2.5.7). Das Perl-Skript kodiert für ein Fenster mit zu definierender Größe, das das Genom

mit einer wählbaren Schrittgröße entlanggleitet. Bei jedem Halt des Fensters wird die genetische Vielfalt des entsprechenden DNA-Bereichs innerhalb des Fensters in Form des H_P-Wertes kalkuliert. Die nötigen Positionsinformationen sowie die Allelfrequenzen in Form von N_i bzw. n_i (s.o.) bezieht das Skript dabei direkt aus der Tab-separierten Textdatei der Variantentabelle. Nach einigen vorausgehenden Analysen (vgl. Kapitel 3.3.3) wurden eine Fenstergröße von 40 kb und eine Schrittgröße von 10 kb festgelegt. Für eine vereinfachte Darstellung in Diagrammen wurden im Anschluss die HP-Werte eines Datensatzes mit den Perl-Skripten sliding_window_parse1.pl sowie sliding_window_parse2.pl chromosomen-weise bzw. genomweit in einer Datei zusammengefasst.

Zum besseren Verständnis des Zusammenhangs zwischen genetischer Diversität und gepoolter Heterozygosität soll im Folgenden der H_P-Wert für zwei vereinfachte Beispielfenster mit nur drei Polymorphismen exemplarisch berechnet werden.

Beispielfenster 1: Es liegen drei polymorphe Positionen innerhalb des Fensters vor. Keine der Varianten verleiht einen Selektionsvorteil.

Tabelle 15: Vereinfachte Variantentabelle eines Beispielfensters ohne Selektion

Position	Varianten	Zahl der *reads*		Abdeckung
		majores Allel	minores Allel	
1	A/G	55	53	108
2	C/T	49	47	96
3	A/C	58	53	111

$$H_P = \frac{2 \times (55 + 49 + 58) \times (53 + 47 + 53)}{((55 + 49 + 58) + (53 + 47 + 53))^2} = 0{,}4997$$

Die genetische Diversität und somit der H_P-Wert des DNA-Abschnitts innerhalb des Beispielfensters 1 sind hoch.

Beispielfenster 2: Es liegen drei polymorphe Positionen innerhalb des Fensters vor. Eine der Varianten (C an Position 2) verleiht ihren Trägern einen Selektionsvorteil. Infolge der genetischen Kopplung werden neutrale Varianten in der Umgebung (A an Position 1 und 2) mitselektiert (*genetic hitchhiking*).

Tabelle 16: Vereinfachte Variantentabelle eines Beispielfensters mit Selektion

Position	Varianten	Zahl der *reads*		Abdeckung
		majores Allel	minores Allel	
1	A/G	92	7	99
2	C/T	101	11	112
3	A/C	98	6	104

$$H_P = \frac{2 \times (92 + 101 + 98) \times (7 + 11 + 6)}{((92 + 101 + 98) + (7 + 11 + 6))^2} = 0{,}1408$$

Die genetische Diversität und somit der H_P-Wert des DNA-Abschnitts innerhalb des Beispielfensters 2 sind erniedrigt.

2.5.9 Z-Transformation

Zur Normalisierung und besseren Visualisierung wurden die H_P-Werte einer Z-Transformation unterzogen. Dabei wurde mithilfe der untenstehenden Formel automatisiert durch das Skript ZH_P_calculator.pl für jedes Fenster ein ZH_P-Wert berechnet.

$$ZH_P = \frac{H_P - \mu H_P}{\sigma H_P}$$

Die Grundlage für die Kalkulation des Mittelwerts μ und der Standardabweichung σ bildeten alle H_P-Werte eines Datensatzes.

2.5.10 Festlegung des Schwellenwertes zur Identifizierung der Kandidatenregionen

Kandidatenregionen, die in der Vergangenheit potenziellen Selektionsereignissen unterlagen, wurden anhand ihrer reduzierten genetischen Diversität bei der Kartierung der Daten der Pool-Sequenzierungen identifiziert. Die genetische Diversität eines genomischen Abschnitts wurde hierbei in Form der gepoolten Heterozygosität (H_P) ermittelt. Wie anhand der Beispielfenster (Tabelle 15 und Tabelle 16) in Kapitel 2.5.8 gezeigt wurde, reflektieren erniedrigte H_P-Werte eine Reduktion der genetischen Diversität im analysierten DNA-Abschnitt. Daher wurde gemäß Qanbari et al. (2012) für jeden der vier Pools ein empirischer Schwellenwert festgelegt, bei dessen Unterschreitung ein Fenster als signifikant und der zugrundeliegende genetische Abschnitt als Kandidatenregion für ein

Selektionsereignis gewertet wurde. Der von Qanbari et al. (2012) entwickelte statistische Test überprüft zu diesem Zwecke, ob die erniedrigten H_P-Werte auch durch Zufall, d. h. in Abwesenheit von Selektion, entstanden sein könnten. Hierfür wurden mittels einer Permutationsmethode die im jeweiligen Datensatz erhobenen Allelfrequenzen gemischt und auf die gegebenen Positionen verteilt. Im Anschluss wurde die gepoolte Heterozygosität für den auf diese Weise zufällig entstandenen Datensatz, wie in Kapitel 2.5.8 beschrieben, kalkuliert und der niedrigste H_P-Wert in einer Textdatei gespeichert. Diese Simulation wurde für jeden der vier Pools 10.000-fach wiederholt und die 10.000 durch Zufall entstandenen H_P-Minima aufgezeichnet. Diese 10.000 niedrigsten H_P-Werte wurden ihrer Größe nach aufsteigend sortiert und der empirische Schwellenwert als der zehntniedrigste Wert dieser Reihe abgelesen. Dieser Wert kommt im analysierten Datensatz nur einmal in 1000 mal durch Zufall zustande, was einem Signifikanzniveau von $p \leq 0{,}001$ entspricht.

Alle Fenster eines Pools, die den ermittelten Schwellenwert unterschritten, wurden in der Kartierung als „Signal“ annotiert und in Form einer .fasta-Datei exportiert. Benachbarte Fenster, die gleichermaßen den Schwellenwert unterschritten, wurden zu einem gemeinsamen Signal zusammengefasst. Mithilfe der BLAST-Funktion der CLC Genomics Workbench 8.0 wurden die den Signalen zugrundeliegenden Kandidatenregionen im Referenzgenom des Pinot Noir Klons PN40024 identifiziert. Als Datenbank für die BLAST-Suche diente die Version 12X.0 des Referenzgenoms, da für diese Version eine Annotation zur Verfügung stand. Sie wurde genau wie die für die Kartierung verwendete Version 12X.2 von der Homepage der „Unité de Recherches en Génomique Info“ (URGI) bezogen. Für die aktuellere Version 12X.2 des Referenzgenoms war zum Zeitpunkt der Analyse keine Annotation vorhanden. Die in den Kandidatenregionen annotierten Gene wurden extrahiert und in Form eine multifasta-Datei sowie als Liste der Gen-IDs gespeichert. Gene, die sich nur zum Teil, beispielweise mit dem 5‘- oder 3‘-Ende, in einer Kandidatenregion befanden, wurden ebenfalls aufgenommen.

2.5.11 GO-Annotation

Die funktionelle Charakterisierung der Kandidatengene erfolgte durch eine GO-Annotation mit dem Online-Tool PLAZA (Proost et al. 2015). Hierbei wurden den Kandidatengenen sogenannte GO-Terms zugeordnet, die Rückschlüsse auf ihre molekulare Funktion und ihre Beteiligung an biologischen Prozessen sowie deren zelluläre Komponenten ermöglichen.

Darüber hinaus wurden durch PLAZA putative Orthologe in *Arabidopsis thaliana* identifiziert und durch die „AnnoMine“-Funktion genauer beschrieben. Durch InterPro- und SignalIP-Plugins wurden konservierte Domänen und Signalpeptide in den abgeleiteten Aminosäuresequenzen der Kandidatengene lokalisiert.

Mithilfe der „*GO Enrichment*“-Funktion von PLAZA wurde überprüft, ob einzelne GO-Terms in den Kandidatengenen im Vergleich zum Gesamtgenom überrepräsentiert sind. Hierzu wurden die Gen-ID-Liste der Kandidatengene dem Arbeitsbereich (*workbench*) von PLAZA hinzugefügt und bei der Analyse das in der Datenbank hinterlegte Referenzgenom von *Vitis vinifera* subsp. *vinifera* als Hintergrundmodell (*background model*) ausgewählt. Zur statistischen Absicherung für multiples Testen wurde eine Bonferroni-Korrektur vorgenommen. GO-Terms mit $p \leq 0{,}01$ wurden als signifikant angereichert angesehen.

3 Ergebnisse

3.1 Charakterisierung der Weinrebe L-17-12-2

Für einen Vergleich der Genome der Wilden Weinrebe *Vitis vinifera* subsp. *sylvestris* und der Edlen Weinrebe *Vitis vinifera* subsp. *vinifera* ist eine eindeutige Zuordnung der sequenzierten Weinreben zu einer der beiden Subspezies unabdingbar. Aufgrund einiger Ergebnisse der vorangegangenen Diplomarbeit bestanden begründete Zweifel an der Klassifizierung der Weinrebe L-17-12-2 als Wilde Weinrebe (Fischer 2012). Daher wurde eine genauere phänotypische und genotypische Charakterisierung der Weinrebe L-17-12-2 vorgenommen.

Als klassisches phänotypisches Unterscheidungsmerkmal zwischen den beiden Subspezies dient das Geschlecht der Weinreben. Daher wurde die Blütenmorphologie kurz nach dem Abwurf der Blütenkäppchen betrachtet (Abbildung 6). Als Vergleichsobjekte wurden eine weibliche und eine männliche Wilde Weinrebe sowie eine zwittrige Kulturrebe (Rebsorte Chardonnay) herangezogen. Alle vier untersuchten Weinreben tragen rispenartige Blüten-

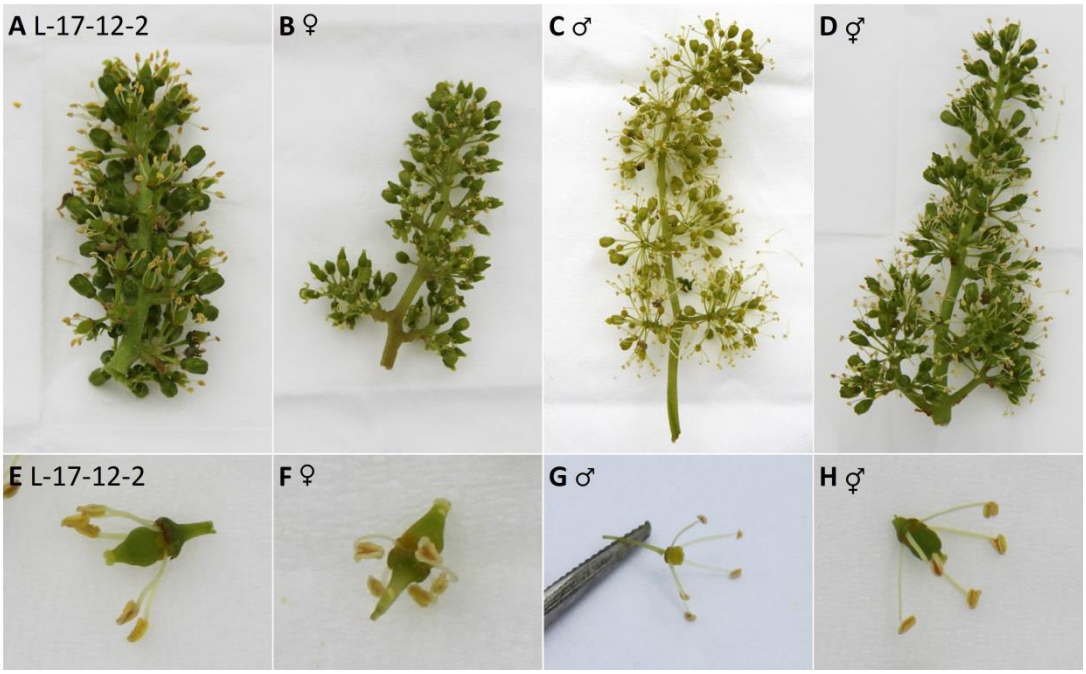

Abbildung 6: Geschlecht der Blüten im Vergleich
Gezeigt sind die Blütenstände (Gescheine) der Weinrebe L-17-12-2 (**A**), einer weiblichen Wilden Weinrebe (**B**), einer männlichen Wilden Weinrebe (**C**) und einer hermaphroditischen Kulturrebe (**D**). Detailaufnahmen zeigen einzelne Blüten der Weinrebe L-17-12-2 (**E**), einer weiblichen Wilden Weinrebe (**F**), einer männlichen Wilden Weinrebe (**G**) und einer hermaphroditischen Kulturrebe (**H**). (Aufnahmen: Sabine Fischer, 12.06.2012)

stände aus über 100 kleinen Einzelblüten (Abbildung 6 A–D). Die Blüten der Weinrebe L-17-12-2 zeigen aufgerichtete Staubblätter, die kreisförmig um ein vollständig ausgebildetes Pistill angeordnet sind (Abbildung 6 E). Die Blüten der weiblichen Wilden Weinrebe besitzen zwar ebenfalls ein voll entwickeltes Pistill bestehend aus Ovarium, Stylus und Stigma; die Filamente der Stamina sind jedoch zurückgebogen (Abbildung 6 F). Bei den Blüten der männlichen Wilden Weinrebe sind die Filamente der Staubblätter aufgerichtet. Das Pistill ist auf ein kleines Ovarium reduziert, Stylus und Stigma sind nicht ausgebildet (Abbildung 6 G). Die hermaphroditischen Blüten der in Abbildung 6 H gezeigten Edlen Weinrebe gleichen dem untersuchten Geschein von L-17-12-2. Sowohl Gynoeceum als auch Androeceum sind voll entwickelt.

Zwar sind die genauen molekularen Mechanismen der Geschlechtsbestimmung in der Weinrebe bisher nicht aufgeklärt, jedoch konnte in den vergangenen Jahren die zugrundeliegende genomische Region auf Chromosom 2 identifiziert und kartiert werden (Picq et al. 2014; Fechter et al. 2012). Daher existieren Marker, die einen Nachweis der verschiedenen Allele des Geschlechtslocus ermöglichen. Mithilfe dieser Marker wurde der entsprechende Locus von L-17-12-2 genauer analysiert und mit diözischen Wilden Weinreben sowie hermaphroditischen Kulturreben verglichen (Abbildung 7). Insgesamt konnten nach gelelektrophoretischer Auftrennung der PCR-Produkte drei Amplifikate detektiert und zwei der drei mittels Sanger-Sequenzierung überprüft werden. BLAST-Analysen ergaben, dass es sich bei dem kleinsten Fragment um das 811 bp lange weibliche F-Allel handelt. Das mittlere PCR-Produkt entspricht dem 879 bp langen M/H-Allel bei dem eine Unterscheidung zwischen der männlichen und der hermaphroditischen Variante nicht möglich ist. Der Größenunterschied zwischen den beiden Allelen ergibt sich durch insgesamt drei Insertions-Deletions-Polymorphismen (Indels), die Größen von 70 bp und zweimal 1 bp aufweisen. Die einzelnen Allele und ihre strukturellen Unterschiede wurden im Rahmen der Dissertation von Birte Ding (2016) genauer analysiert. Bei der obersten Bande handelt es sich höchstwahrscheinlich um ein Hybrid aus den anderen beiden Banden. Sequenzhomologe Bereiche der zwei Amplifikate lagern sich während der Abkühlphase des PCR-Zyklus zusammen, die Insertion bildet eine Schleife aus. Dies wurde mittels Verwendung eines denaturierenden Polyacrylamid-Gels bestätigt (Birte Ding, pers. Mitteilung). Anhand der in Abbildung 7 gezeigten gelelektrophoretischen Auftrennung der Marker-PCR lassen sich folgende Genotypen für die einzelnen Weinreben ableiten: L-17-12-2 ist genau wie die männliche Wilde Weinrebe und eine der beiden zwittrigen

Kulturreben heterozygot und trägt sowohl das F-Allel als auch das MH-Allel. Die weibliche Wilde Weinrebe trägt homozygot das F-Allel. Die verbleibende zweite Edle Weinrebe ist homozygot für das MH-Allel.

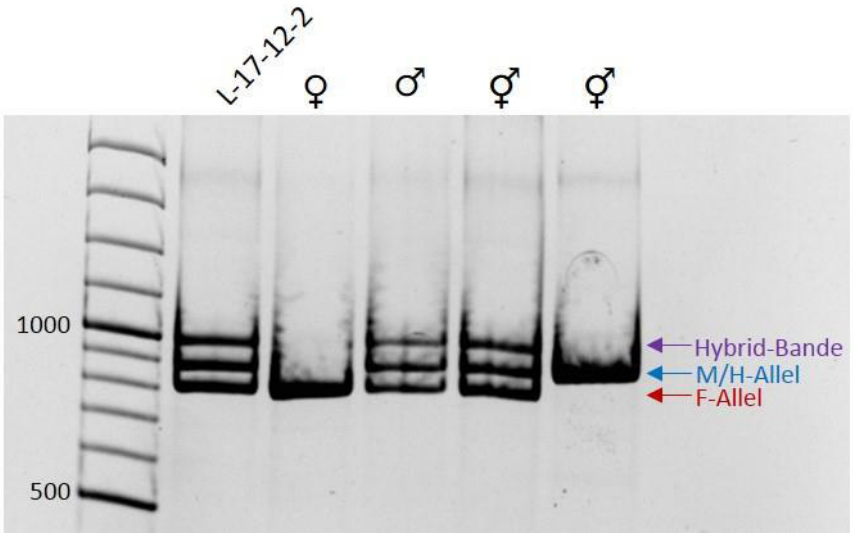

Abbildung 7: Analyse des Geschlechtslocus von L-17-12-2
Das Bild zeigt die gelelektrophoretische Auftrennung der Produkte einer PCR auf die geschlechtsgekoppelten Marker von Fechter et al. (2012). Die mit rotem Pfeil markierten Banden entsprechen dem 811 bp großen F-Allel, bei den mit blauem Pfeil markierten Banden handelt es sich um das 879 bp große M/H-Allel. Dies wurde per Sanger-Sequenzierung bestätigt. In Fällen, bei denen sowohl ein F-Allel als auch ein M/H-Allel amplifiziert wurde, bildet sich eine zusätzliche Hybrid-Bande aus beiden Fragmenten (lila Pfeil). Von links nach rechts sind die PCR-Produkte von L-17-12-2, einer weiblichen Wilden Weinrebe, einer männlichen Wilden Weinrebe, der hermaphroditischen Kulturrebe Weißer Heunisch und der hermaphroditischen Kulturrebe Chardonnay aufgetragen.

Als weiteres Merkmal wurde die Beerenfarbe und die der Färbung zugrundeliegende genomische Region der Weinrebe L-17-12-2 analysiert. Die Rebstöcke von L-17-12-2 tragen im Herbst Traubenbündel, deren Beeren bei der Reife nicht den für Wilde Weinreben typischen schwarz-blauen, sondern einen rosa Farbton aufweisen (Abbildung 8 A). Verantwortlich für die Färbung des Beerenexokarps sind zwei Transkriptionsfaktoren der MYB-Familie, die die Anthocyanbiosynthese regulieren. Die entsprechenden Gene – *VvMybA1* und *VvMybA2* – befinden sich geclustert mit weiteren Mitgliedern dieser Genfamilie auf Chromosom 2. Sie wurden auf genomischer Ebene mittels PCR und Sanger-Sequenzierung untersucht (Abbildung 8 B–E). Ein Allel des *VvMybA1*-Gens weist im Promotorbereich 181 bp stromaufwärts des Startkodons ein *Gypsy*-LTR-Retrotransposon auf. Alignments mit Datenbankeinträgen zeigten, dass es sich dabei um das sogenannte weiße Allel des Farblocus handelt, bei dem die Integration des *Gypsy*-LTR-Retrotransposons

Gret1 (*grapevine retrotransposon 1*) im Promotorbereich die Expression des *VvMybA1*-Gens verhindert. Die zweite Kopie des *VvMybA1*-Gens auf dem homologen Chromosom weist einen intakten Promotor ohne Retrotransposon auf. Der Promotor dieser Kopie wurde im Rahmen der Bachelorarbeit von Sarah Dietzen (2012) genauer analysiert. Dabei stellte sich heraus, dass sich die Promotorsequenz von L-17-12-2 durch zahlreiche Polymorphismen, darunter ein 44 bp langes Indel, von den Promotoren klassischer roter Rebsorten unterscheidet. Für das *VvMybA2*-Gen konnte gezeigt werden, dass der offene Leserahmen durch eine homozygote Deletion zweier Nukleotide im dritten Exon und der daraus folgenden Leserasterverschiebung zerstört ist (Abbildung 8 E).

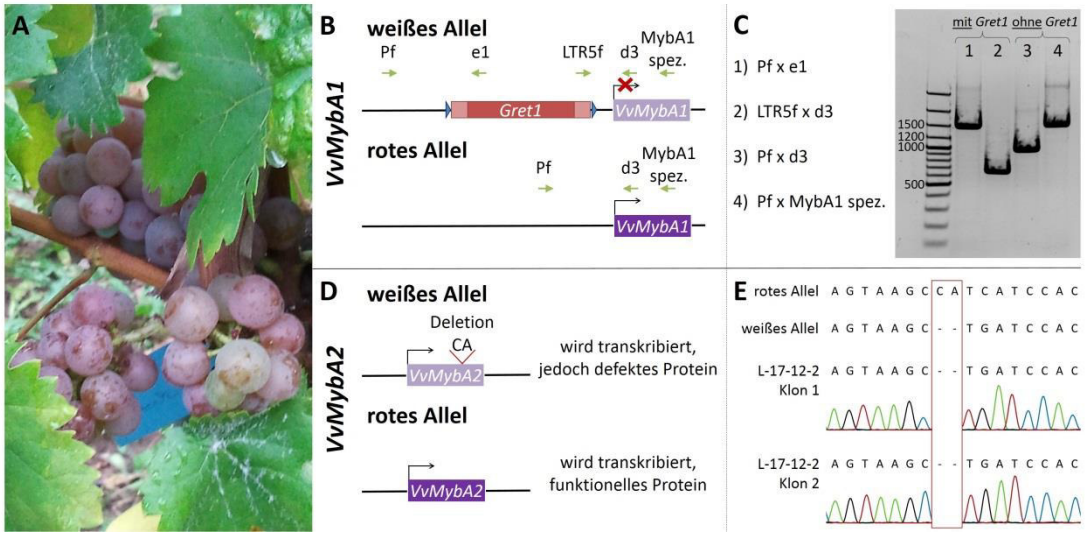

Abbildung 8: Analyse des Farblocus von L-17-12-2

A zeigt die Traubenbündel von L-17-12-2 kurz vor der Reife (Aufnahme: Sabine Fischer, 11.09.2012). **B** liefert einen schematischen Überblick des *VvMybA1*-Locus inklusive Lage der verwendeten Primer. Im weißen Allel ist das *VvMybA1*-Gen durch die Insertion von *Gret1* im Promotorbereich inaktiviert. Das rote Allel besitzt eine intakte *VvMybA1*-Variante, von der ein funktioneller Transkriptionsfaktor gebildet werden kann. **C** zeigt die gelelektrophoretische Auftrennung der PCR-Produkte von *VvMybA1*. Mit allen vier Primerpaaren wurden Fragmente der erwarteten Größe amplifiziert. **D** liefert einen schematischen Überblick des *VvMybA2*-Locus. Im weißen Allel verursacht die Deletion zweier Nukleotide eine *frameshift*-Mutation. Das Proteinprodukt ist, sofern es überhaupt gebildet wird, nicht funktionell. Im roten Allel ist das Gen intakt und das abgeleitete Protein funktionell. **E** zeigt einen Ausschnitt der Sanger-Sequenzierung des *VvMybA2*-Gens von L-17-12-2. In den beiden oberen Zeilen sind das rote und das weiße Allel als Referenz angegeben. Darunter sind exemplarisch die Elektropherogramme zweier Klone gezeigt. In beiden sind zwei Nukleotide (CA) deletiert. Insgesamt wurden 20 Klone analysiert. Die Deletion ist in allen Klonen vorhanden.

Um die Identität der Weinrebe L-17-12-2 abschließend zu klären, wurde ein genetischer Fingerabdruck angefertigt. Dies geschah mithilfe von neun hochauflösenden SSR-Markern. Die vermessenen Allelgrößen der einzelnen Loci sind in Tabelle 17 aufgeführt. Für drei der

neun untersuchten Loci wurden Allelgrößen gemessen, die in dieser Form nicht bei *Vitis vinifera* subsp. *sylvestris* existieren. Es ist daher auszuschließen, dass es sich bei L-17-12-2 um eine reine Wilde Weinrebe handelt. Die identifizierten Allelgrößen deuten darauf hin, dass eine Hybride aus *Vitis vinifera* subsp. *vinifera* und *Vitis vinifera* subsp. *sylvestris* vorliegt.

Tabelle 17: SSR-Profil von L-17-12-2

Locus	Allel 1	Allel 2
VVMD5	234 bp	240 bp
VVS2	151 bp	151 bp
VVMD7	243 bp[1]	243 bp[1]
MD27/ZAG47	190 bp	190 bp
VrZAG62	188 bp[1]	188 bp[1]
VrZAG79	245 bp	245 bp
VVMD32	272 bp	272 bp
VVMD28	218 bp[1]	240 bp[1]
VVMD25	249 bp	249 bp

[1]Rot markierte Allele kommen in dieser Größe nicht bei *Vitis vinifera* subsp. *sylvestris* vor.

3.2 Komparative Genomanalysen der Edlen Weinrebe und der Wilden Weinrebe

3.2.1 Genomische Illumina-Daten von Weinreben

Zur Charakterisierung des Genoms der Wilden Weinrebe und für Vergleiche mit dem Genom der Edlen Weinrebe wurden insgesamt acht genomische *paired-end* DNA-Bibliotheken im Hochdurchsatzverfahren sequenziert. Ausgangsmaterial für vier der acht Bibliotheken war die genomische DNA einzelner Individuen, darunter zwei Wilde Weinreben, eine Edle Weinrebe und die bereits besprochene Hybridrebe L-17-12-2. Bei den verbleibenden vier Bibliotheken handelt es sich um sogenannte Pools, bei denen die genomische DNA von 10–15 Individuen vor der Erstellung der Bibliotheken vereint wurde. Die Individuen eines Pools stammen hierbei aus einer gemeinsamen Ursprungsregion. Insgesamt wurden mit dem Illumina HiSeq 2000 sowie dem Nachfolgemodell HiSeq 2500 1.356.019.644 Sequenzen (*reads*) generiert. Die Länge dieser Rohdaten betrug abhängig von der verwendeten Chemie 97–108 bp. Demzufolge liefern die Sequenzierungen eine Gesamtinformation von rund 140 Gb. Positionen mit undefinierten Basen (N, etwa 0,05 %) wurden gemeinsam mit verbleibenden Adaptersequenzen sowie Bereichen schlechter Qualität in einem ersten Filter-Schritt beseitigt, um derartige Sequenzen nicht mit in die

Folgeanwendungen zu übernehmen. Zudem wurde am 5‘-Ende eines jeden *reads* pauschal 1 bp und am 3‘-Ende 5 bp entfernt. Sequenzen, die im Anschluss kürzer als 30 bp waren, wurden verworfen. Bei der Aufarbeitung der Rohdaten wurden insgesamt 17,41 % der Nukleotide entfernt, so dass 115 Gb Sequenzinformation in Form der gefilterten Daten für die weiteren Analysen zur Verfügung standen. Tabelle 18 vergleicht die Rohdaten (R) und die gefilterten Daten (G) hinsichtlich Umfang, Sequenzlänge, Basenzusammensetzung und Qualität.

Tabelle 18: Statistische Auswertung der genomischen Illumina-Sequenzdaten

Bibliothek			Sequenzen	Nukleotide	Ø Länge	% GC	% N	Ø Phred
VV	Weißer Heunisch	R	150.642.258	15,21 Gb	101	37,31	0,08	35,06
		G	138.764.543	12,25 Gb	88,27	36,43	0	37,81
Hyb	L-17-12-2	R	171.716.108	17,34 Gb	101	36,78	0,13	34,86
		G	160.481.019	14,00 Gb	87,25	36,02	0	37,64
VS	Hördt29	R	200.258.978	20,23 Gb	101	33,87	0,02	35,16
		G	186.642.479	16,23 Gb	86,98	33,50	0	37,46
VS	Kaukasus	R	19.611.868	1,96 Gb	97/101	35,01	0,05	34,10
		G	17.528.827	1,50 Gb	85,38	34,74	0	37,54
VS	Pool Ketsch	R	234.337.684	24,40 Gb	101/108	35,41	0,02	35,42
		G	225.795.243	20,64 Gb	91,43	34,97	0	37,71
VS	Pool Pisa	R	141.573.206	14,70 Gb	101/108	35,79	0,04	35,15
		G	137.312.853	12,49 Gb	90,95	35,18	0	37,69
VS	Pool Frankreich	R	176.623.184	18,62 Gb	101/108	34,92	0,03	34,84
		G	167.686.496	15,43 Gb	91,99	34,44	0	37,59
VS	Pool Spanien	R	261.256.358	27,28 Gb	101/108	35,69	0,04	35,04
		G	251.036.846	22,88 Gb	91,13	35,10	0	37,59

verwendete Abkürzungen: VV: *Vitis vinifera* subsp. *vinifera*, VS: *Vitis vinifera* subsp. *sylvestris*, Hyb: Hybridrebe, R: Rohdaten, G: gefilterte Daten

Die Sequenzlänge der gefilterten Daten beträgt im Durchschnitt 89,8 bp. Die Spannweite reicht hierbei von 30–102 bp. Der GC-Gehalt sinkt in allen Bibliotheken im Vergleich zu den Rohdaten geringfügig. Die Bibliotheken der Wilden Weinreben weisen allgemein einen niedrigeren GC-Gehalt als die der Edlen Weinrebe und der Hybridrebe auf. Besonders hervorzuheben ist der Datensatz der Wilden Weinrebe Hördt29, dessen GC-Gehalt mit 33,50 % fast 3 % unter dem der Edlen Weinrebe Weißer Heunisch liegt. Es sind, wie gefordert, keine undefinierten Basen mehr in den Sequenzen vorhanden. Die Qualitätswerte haben sich im Zuge der Aufarbeitung verbessert; sie liegen bei den Rohdaten im Rahmen von 34–35, bei den gefilterten Sequenzdaten werden hingegen durchschnittliche Phred-Werte von über 37 erreicht. Da die Aussagekraft eines durchschnittlichen Qualitätswertes für einen gesamten Datensatz begrenzt ist, wurden

darüber hinaus mit dem Programm FastQC Qualitätsdiagramme erstellt in denen die durchschnittlichen Phred-Werte Positions-spezifisch entlang der Sequenz abgebildet werden (Abbildung 9). Es zeigt sich, dass bereits die Rohdaten eine akzeptable Qualität aufweisen. Ab Position 65 lässt sich jedoch anhand der Standardabweichung erkennen, dass sich vermehrt Basen im mittleren (orangenen), ab Position 75 sogar im schlechten (roten) Qualitätsbereich befinden (Abbildung 9 A). Diese erhöhten Fehlerraten konnten durch die beschriebenen Filtermaßnahmen verringert werden, so dass die gefilterten Daten ein durchweg sehr gutes (grünes) Qualitätsbild zeigen (Abbildung 9 B).

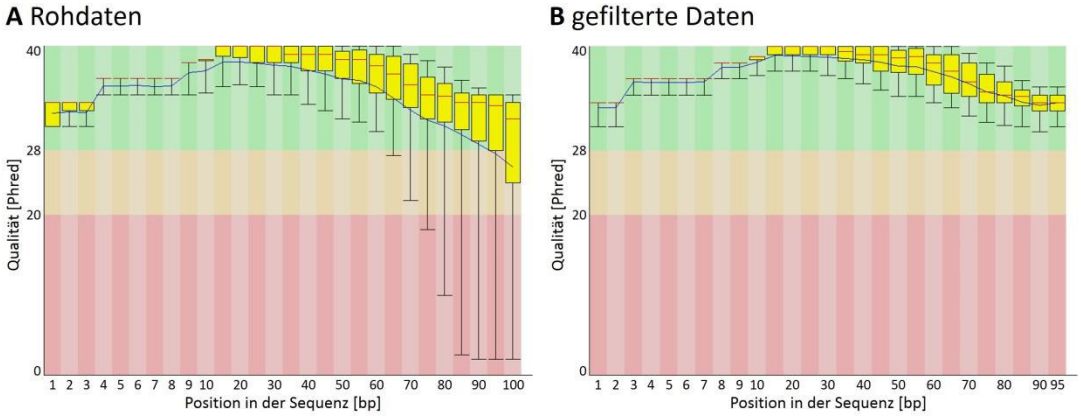

Abbildung 9: Qualitätskontrolle von ungefilterten Rohdaten und gefilterten Daten der Illumina-Sequenzierung

Die Diagramme zeigen exemplarisch eine Qualitätsanalyse des Datensatzes *Vitis vinifera* subsp. *vinifera* var. Weißer Heunisch. Auf der X-Achse ist die Position in der Sequenz, auf der Y-Achse der zugehörige durchschnittliche Qualitätswert der jeweiligen Position angegeben. Grüne Bereiche weisen eine sehr gute Qualität mit Phred-Werten über 28 auf. Orange kennzeichnet Bereiche mittlerer Qualität mit Phred-Werten von 20–28. Schlechte Phred-Werte unter 20 sind rot markiert. **A** FastQC-Diagramm der Illumina-Rohdaten. **B** FastQC-Diagramm der gefilterten Daten.

3.2.2 Abdeckung des Weinrebengenoms

Um zu beurteilen, ob ausreichend Datenmaterial für die geplanten komparativen Genomanalysen vorhanden ist, wurde die Abdeckung (*coverage*) des Genoms der Weinrebe durch die Illumina-Sequenzdaten der einzelnen Bibliotheken abgeschätzt. Hierzu wurde zunächst eine rein rechnerische Abdeckung des Genoms ermittelt, indem die Anzahl der Nukleotide eines Datensatzes durch die Länge des Rebengenoms dividiert wurde (Tabelle 19 linke Spalte). Diese einfache Schätzung kann nur als erster Anhaltspunkt dienen, da sie vernachlässigt, dass nicht alle *reads* im Genom kartieren und die Verteilung der Sequenzen in der Regel nicht homogen ist. Aus diesen Gründen wurde eine alternative Berechnung der

durchschnittlichen Genomabdeckung anhand einer Kartierung (*mapping*) der Illumina-Sequenzdaten gegen das Referenzgenom vorgenommen (Tabelle 19 rechte Spalte). Diese Herangehensweise hat den Vorteil, dass nur Daten, die tatsächlich im Referenzgenom kartieren und somit an dieser Stelle eine Sequenzinformation liefern, in die Berechnung einfließen. Bei den Sequenzierungen der vier Einzelindividuen wird das Genom je nach Datensatz zwischen 22- und 30-fach abgedeckt. Nur die Bibliothek der Wilden Weinrebe aus dem Kaukasus liefert eine deutlich geringere Abdeckung des Genoms. Bei den Pool-Sequenzierungen wird das Genom der Weinrebe 18- bis 33-fach abgedeckt.

Tabelle 19: Abschätzung der Abdeckung des Genoms durch die Illumina-Sequenzdaten

Bibliothek		**rechnerische Abdeckung des Genoms „x“**	**durchschnittliche Genomabdeckung im *Mapping***
VV	Weißer Heunisch	25,19x	22,40
Hyb	L-17-12-2	28,80x	26,51
VS	Hördt29	33,39x	30,81
VS	Kaukasus	3,08x	2,60
VS	Pool Ketsch	42,46x	31,00
VS	Pool Pisa	25,69x	18,72
VS	Pool Frankreich	31,73x	22,75
VS	Pool Spanien	47,05x	33,92

Die Differenz zwischen den Ergebnissen der beiden Berechnungsformen deutet darauf hin, dass ein nicht vernachlässigbarer Anteil der Daten nicht im Referenzgenom kartiert.

Pflanzen wie die Weinrebe besitzen neben der Kern-DNA mit dem Erbgut der Chloroplasten und der Mitochondrien zwei weitere Genome in ihren Zellen. Da diese beiden Genome genau wie das Kerngenom aus DNA bestehen und die entsprechenden Organellen bei der DNA-Extraktion nicht durch spezielle Isolationsschritte entfernt wurden, ist davon auszugehen, dass ein Teil der *reads* von den Genomen der Chloroplasten und Mitochondrien stammt und daher nicht im Referenzgenom kartiert. Um den Anteil solcher Sequenzen zu ermitteln, wurden die Datensätze mit dem Chloroplasten- und dem Mitochondrien-Genom der Weinrebe verglichen (Tabelle 20).

Tabelle 20: Anteil der Chloroplasten- und Mitochondrien-*reads* in den Bibliotheken

Bibliothek		Chloroplasten-*reads*	Mitochondrien-*reads*	Σ
VV	Weißer Heunisch	9,49 %	7,42 %	16,91 %
Hyb	L-17-12-2	9,53 %	9,01 %	18,54 %
VS	Hördt29	9,83 %	6,55 %	16,38 %
VS	Kaukasus	3,99 %	2,34 %	6,33 %
VS	Pool Ketsch	10,23 %	6,80 %	17,03 %
VS	Pool Pisa	10,83 %	6,52 %	17,35 %
VS	Pool Frankreich	12,78 %	8,10 %	20,88 %
VS	Pool Spanien	9,72 %	6,72 %	16,44 %

Der Anteil der Chloroplasten-*reads* beläuft sich auf 9,49 bis 12,78 %, eine Ausnahme stellt der Datensatz Kaukasus dar, hier sind mit 3,99 % deutlich weniger Sequenzen plastidiärer Herkunft. Der mitochondriale Anteil ist mit 6,52 bis 9,01 % vergleichsweise geringer. Auch hier weicht die Bibliothek der Wilden Weinrebe aus dem Kaukasus deutlich von den anderen Datensätzen ab, der mitochondriale *read*-Anteil liegt bei 2,34 %. Zusammengefasst sind bis zu 20 % der Illumina-Sequenzen nicht genomischen Charakters.

3.2.3 *De novo* Assemblierung

Die gefilterten Illumina-Sequenzdaten wurden mit der CLC Assembly Cell assembliert und auf diese Weise zu größeren zusammenhängenden Sequenzfragmenten zusammengesetzt (Tabelle 21). Bei der *de novo* Assemblierung der genomischen Daten wurden für die Einzelsequenzierungen der beiden Wilden Weinreben sowie der Hybridrebe ca. 130.000 solcher Sequenzfragmente, die im Weiteren *contigs* genannt werden, generiert. Für die Pool-Sequenzierungen und die Einzelsequenzierung der Kulturrebe Weißer Heunisch wurden deutlich mehr, nämlich über 200.000 *contigs* erzeugt. Fasst man alle *contigs* eines Datensatzes zusammen, ergeben sich Gesamtlängen der Assemblierungen, die von 299,86 bis 347,61 Mb reichen. Der Datensatz Kaukasus stellt erneut eine Ausnahme dar, hier wird nur eine Gesamtlänge von knapp 50 Mb erreicht.

Tabelle 21: *De novo* Assemblierung der genomischen Illumina-Daten

Datensatz		*contig*-Zahl	Gesamtlänge	Ø *contig*-Länge	N50	% GC
VV	Weißer Heunisch	202.818	345,03 Mb	1.701 bp	4.805 bp	33,80
Hyb	L-17-12-2	131.643	341,04 Mb	2.591 bp	8.197 bp	33,96
VS	Hördt29	134.305	343,31 Mb	2.556 bp	8.566 bp	33,44
VS	Kaukasus	131.925	49,28 Mb	374 bp	373 bp	31,96
VS	Pool Ketsch	205.493	341,94 Mb	1.664 bp	5.187 bp	33,65
VS	Pool Pisa	202.964	299,86 Mb	1.477 bp	3.922 bp	33,64
VS	Pool Frankreich	209.895	319,00 Mb	1.520 bp	4.245 bp	33,42
VS	Pool Spanien	243.552	347,61 Mb	1.427 bp	4.542 bp	33,53

Die Qualität einer Assemblierung lässt sich anhand der durchschnittlichen Länge ihrer *contigs* oder auch anhand des N50-Wertes beurteilen. Der N50-Wert entspricht der Länge des kürzesten *contigs* in einer Gruppe, die die größten *contigs* einer Assemblierung enthält und deren kombinierte Länge mindestens 50 % der Gesamtlänge der Assemblierung beträgt (Miller et al. 2010). Die besten Werte erreichen die Einzelsequenzierung der Hybridrebe und der Wilden Weinrebe Hördt29 mit einer durchschnittlichen *contig*-Länge von über 2.500 bp und einem N50-Wert größer 8.000 bp. Die Assemblierungen der Pool-Sequenzierungen sowie der Kulturrebe Weißer Heunisch, die zuvor durch die größere Anzahl erzeugter c*ontigs* hervorstachen, sind von deutlich schlechterer Qualität. Im Fall der Wilden Weinrebe aus dem Kaukasus ist die Assemblierung mit einer durchschnittlichen c*ontig*-Länge und einem N50-Wert von jeweils unter 400 bp nicht weiter verwendbar. Der GC-Gehalt der assemblierten Daten hat im Vergleich zu den gefilterten Daten abgenommen, er liegt bei 31,96 bis 33,96 %.

3.2.3 Transposable Elemente im Weinrebengenom

Transposable Elemente und mit ihnen verwandte Sequenzen nehmen bei vielen Pflanzen einen Großteil der Genome ein. Obwohl ihre Funktion oft ungeklärt ist, existieren Beispiele in denen sie eine wichtige Rolle bei der Evolution und Domestizierung von Kulturpflanzen spielen (Studer et al. 2011; Otto et al. 2014). Daher wurde hier der Transposongehalt in den Genomen der sequenzierten Kultur- und Wildreben untersucht. Zum Vergleich wurden die Illumina-Sequenzdaten von drei weiteren Weinreben hinzugezogen. Darunter befanden sich mit der Rotweinsorte Tannat und der Tafeltraube Sultanina (auch „Sultana“ oder „*Thompson seedless*“ genannt) zwei Vertreter der Edlen Weinrebe *Vitis vinifera* subsp. *vinifera* sowie die amerikanische *Vitis*-Spezies *Vitis rotundifolia*. Diese Daten wurden über das *Sequence Read Archive* (SRA) von NCBI bezogen (Akzessionsnummern: SRR863618, SRR863595, SRR924200, SRR094945). Für die Bestimmung des Transposongehalts der Genome wurde angenommen, dass die Häufigkeit eines DNA-Elements direkt mit der Anzahl der zugehörigen Sequenzen in einem Illumina-Datensatz korreliert. Anders ausgedrückt, wenn ein transposables Element in einem Genom doppelt so oft vorhanden ist wie in einem anderen Genom, so existieren im Datensatz des ersten Genoms auch doppelt so viele *reads* dieses Transposons. Die ermittelten Werte wurden auf die Länge der transposablen Elemente und die Größe des jeweiligen Datensatzes normalisiert, um die Vergleichbarkeit über verschiedene transposable Elemente und Genome hinweg zu

gewährleisten. Mithilfe von Referenzwerten von drei *single copy*-Genen der Weinrebe wurde die absolute Kopienzahl der transposablen Elemente in den verschiedenen analysierten Genomen abgeschätzt. Aufgrund der geringen Datensatzgröße der Wilden Weinrebe aus dem Kaukasus war keines der gewählten *single copy*-Gene ausreichend mit Sequenzen abgedeckt, um den benötigten Referenzwert zu berechnen. Daher konnten für die transposablen Elemente dieser Probe keine Kopienzahlen ermittelt werden.

Transposable Elemente lassen sich basierend auf ihrem Transpositionsmechanismus in zwei große Gruppen untergliedern (Finnegan 1989). Klasse I Transposons verwenden bei der Transposition ein RNA-Molekül als Zwischenstufe und werden auch als Retrotransposons bezeichnet. Bei den Klasse II Transposons erfolgt die Transposition hingegen über ein DNA-Molekül, dementsprechend spricht man in diesem Fall von DNA-Transposons. In allen untersuchten Genomen der Kultur- und Wildreben übersteigt die Kopienzahl der Klasse I Transposons die der Klasse II Transposons um ein Vielfaches (Abbildung 10). Die größte Zahl von sowohl Klasse I als auch Klasse II Transposons wurde im Pool Pisa identifiziert, hier finden sich mehr als 22.400 Kopien von Retrotransposons und ca. 3.800 DNA-Transposon-Kopien.

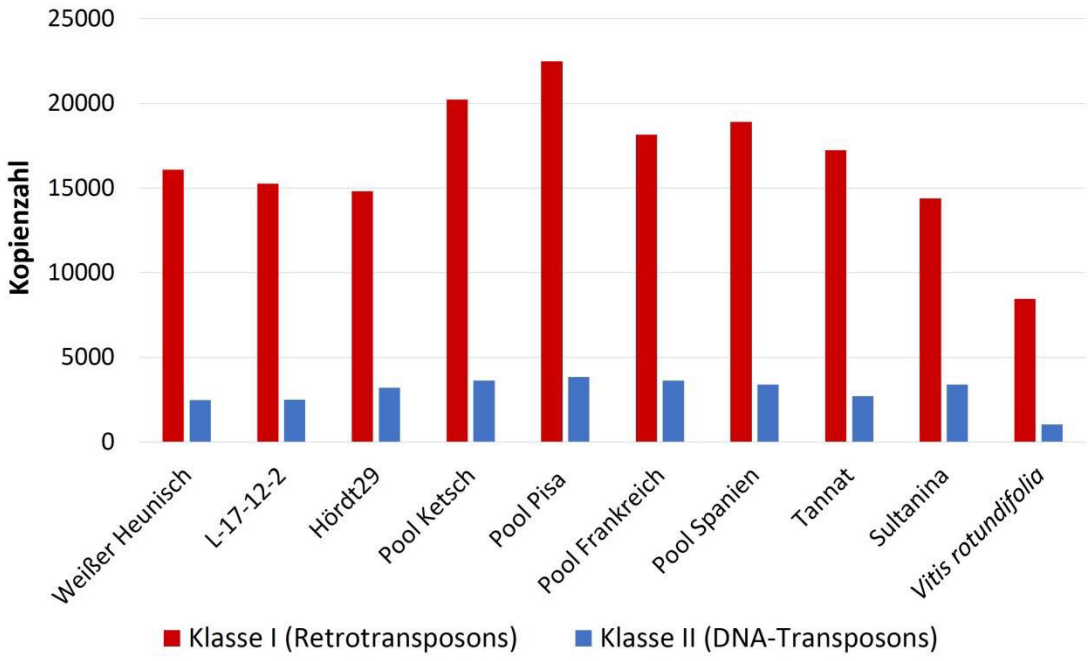

Abbildung 10: Kopienzahl der Klasse I und Klasse II Transposons in den verschiedenen Rebengenomen
Das Diagramm zeigt die Zahl der Kopien pro Genom der identifizierten Klasse I und Klasse II Transposons. Zu erkennen ist ein deutlicher Überschuss an Klasse I Transposons in allen analysierten Rebengenomen. Die größten Kopienzahlen wurden in den Pool-Sequenzierungen nachgewiesen.

Allgemein wurden in den Pool-Sequenzierungen höhere Kopienzahlen als in den Sequenzierungen von Einzelindividuen festgestellt. Die geringste Anzahl transposabler Elemente wurde in *Vitis rotundifolia* gemessen, insbesondere die DNA-Transposons liegen im Vergleich zu den anderen analysierten Genomen in stark reduzierter Kopienzahl vor.

Die transposablen Elemente beider Klassen lassen sich in Familien gruppieren, die im Weiteren untersucht wurden (Abbildung 11). Unter den Retrotransposons dominieren die Elemente der *Copia*- und *Gypsy*-Familie mit Kopienzahlen von 4.000 bis mehr als 12.000 Elementen pro Genom. Beide Familien zeichnen sich durch lange, terminale Wiederholungseinheiten (*long terminal repeats*, LTR) an beiden Enden aus, weshalb sie zur Überfamilie der LTR-Retrotransposons zusammengefasst werden. Ein Vergleich beider Familien untereinander zeigt, dass in den einzelnen analysierten Genomen jeweils mehr *Gypsy*-Transposons als Elemente der *Copia*-Familie vorhanden sind. In der Tafeltraube Sultanina und der amerikanischen *Vitis*-Spezies *Vitis rotundifolia* fällt dieser Unterschied

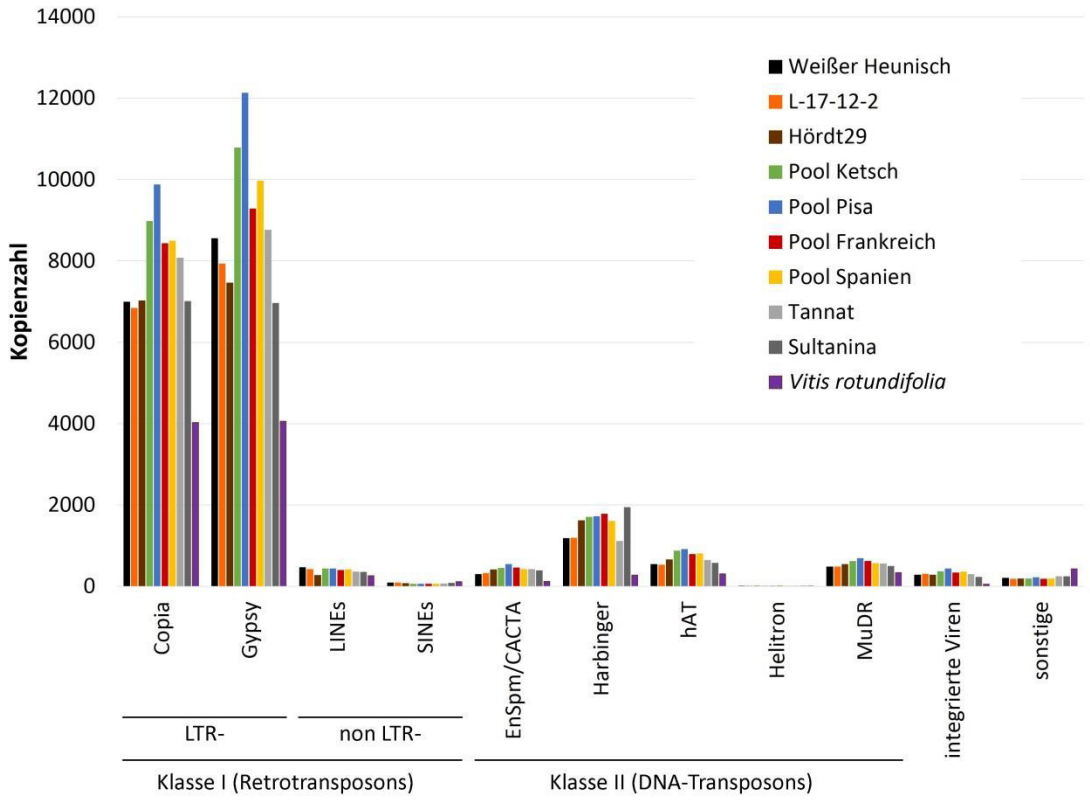

Abbildung 11: Kopienzahl der Transposonfamilien und anderer verwandter Elemente in den verschiedenen Rebengenomen
Das Diagramm zeigt die Kopienzahlen der unterschiedlichen Transposonfamilien. Die Klasse I Retrotransposons dominieren in Form der *Gypsy*- und *Copia*-Elemente deutlich in der Kopienzahl in allen analysierten Rebengenomen.

jedoch deutlich geringer aus als in den verbleibenden untersuchten Genomen. Neben den angesprochenen LTR-Retrotransposons wurden zwei weitere Familien der Klasse I Transposons in den sequenzierten Genomen der Kultur- und Wildreben identifiziert, die jedoch keine flankierenden Wiederholungseinheiten tragen und daher als *non* LTR-Transposons zusammengefasst werden. Dabei handelt es sich um die Familien der LINEs und der SINEs, die in den untersuchten Genomen mit deutlich geringeren Kopienzahlen als *Copia*- und *Gypsy*-Elemente vertreten sind. Im Falle der LINEs reicht die Kopienzahl von 262 bis 459. Von den SINE-Elementen wurden sogar nur ca. 50 bis 110 Kopien pro Genom identifiziert. Aus der Klasse der DNA-Transposons wurden mit EnSpm/CACTA, Harbinger, hAT, Helitron und MuDR insgesamt fünf verschiedene Familien analysiert. Auch sie treten signifikant seltener auf als die *Copia*- und *Gypsy*-Elemente. Die am weitesten verbreiteten Klasse I Transposons sind die Mitglieder der Harbinger-Familie mit bis zu knapp 2.000 Kopien pro Genom. Auffällig ist hier eine höhere Kopienzahl insbesondere in den Wilden Weinreben im Vergleich zu den Kulturreben und der Hybridrebe. In den Wilden Weinreben ließen sich 1.600 bis 1.774 Kopien der Harbinger-Transposons identifizieren. In den Kulturreben bzw. der Hybridrebe wurden hingegen nie mehr als ca. 1.200 Kopien gemessen. Eine Ausnahme stellt hierbei nur die Tafeltraube Sultanina dar, deren Werte sogar die der Wilden Weinrebe übertreffen. Die verbleibenden vier Familien der DNA-Transposons zeigen Kopienzahlen kleiner 1.000. Im Falle der Helitrons konnte in allen zehn untersuchten Genomen nur jeweils eine Helitron-Kopie identifiziert werden. Neben den eigentlichen transposablen Elementen wurden darüber hinaus Transposon-verwandte Sequenzen wie integrierte Virusfragmente sowie Transposons, die sich in keine der beiden Klassen einordnen lassen, analysiert. Keines dieser Elemente ist mehr als 500-mal im Genom vorhanden. Bemerkenswert ist jedoch das unklassifizierte Element *Simple-1_VVi*, das in *Vitis rotundifolia* im Vergleich zu den anderen analysierten Genomen eine deutliche Erhöhung der Kopienzahl auf fast 400 Kopien zeigt.

Von der ermittelten Kopienzahl wurde auf die Gesamtlänge der transposablen Elemente in den analysierten Genomen geschlossen. Hierfür wurde die Kopienzahl mit der durchschnittlichen Länge des entsprechenden transposablen Elements multipliziert. Auf diese Art und Weise wurden Gesamtlängen ermittelt, die von 32,75 Mb in der amerikanischen *Vitis*-Spezies *Vitis rotundifolia* bis 84,77 Mb in der Pool-Sequenzierung der Wilden Weinreben aus Pisa reichen (Tabelle 22). Die bereits beschriebene höhere Kopienzahl an transposablen Elementen in den Pool-Sequenzierungen im Vergleich zu den

Einzelindividuen spiegelt sich erwartungsgemäß auch in einer größeren Gesamtlänge wider. Betrachtet man ausschließlich die sequenzierten Einzelindividuen, so führen hinsichtlich der Gesamtlänge der identifizierten transposablen Elemente die beiden für die Weinproduktion angebauten Edelreben Tannat und Weißer Heunisch vor der Hybridrebe L-17-12-2 gefolgt von der Wilden Weinrebe Hördt29 sowie der Tafeltraube Sultanina. *Vitis rotundifolia* befindet sich mit großem Abstand am Ende dieser Rangfolge.

Eine Berechnung des prozentualen Anteils transposabler Elemente am Gesamtgenom gestaltete sich schwierig, da die Genomgröße der verschiedenen Weinreben unbekannt und auf Basis der vorhandenen Illumina-Daten nicht einfach zu berechnen ist. Daher wurde mangels Alternativen für alle betrachteten Weinreben die Größe des von Jaillon et al. (2007) publizierten Referenzgenoms herangezogen. Der Anteil der identifizierten transposablen Elemente und mit ihnen verwandten Sequenzen am Gesamtgenom reichte von 6,74 % in der amerikanischen *Vitis*-Spezies *Vitis rotundifolia* bis zu 17,44 % im Pool der Wilden Weinreben aus Pisa.

Tabelle 22: Gesamtlänge und genomischer Anteil[1] der identifizierten transposablen Elemente

Genom	Gesamtlänge transposabler Elemente	Prozentualer Anteil[1] transposabler Elemente am Genom
Weißer Heunisch	61,01 Mb	12,55 %
L-17-12-2	57,81 Mb	11,89 %
Hördt29	54,78 Mb	11,27 %
Pool Ketsch	76,55 Mb	15,74 %
Pool Pisa	84,77 Mb	17,44 %
Pool Frankreich	68,73 Mb	14,14 %
Pool Spanien	70,93 Mb	14,59 %
Tannat	64,22 Mb	13,21 %
Sultanina	51,69 Mb	10,63 %
Vitis rotundifolia	32,75 Mb	6,74 %

[1] Der hier angegebene prozentuale genomische Anteil bezieht sich auf das publizierte Referenzgenom von Jaillon et al. (2007).

Aufgrund der in Abbildung 11 dargestellten Unterschiede in der Kopienzahl zwischen den verschiedenen Weinreben eignen sich die identifizierten transposablen Elemente als molekulare Marker für die Weinrebenforschung und –züchtung. Daher wurden im Rahmen der Bachelorarbeit von Michel Seiwert (2014) verschiedene publizierte Transposon-basierte Marker getestet und molekulare Fingerabdrücke einer Auswahl von Kultur- und Wildreben angefertigt. Die hierzu eingesetzen Methoden IRAP (*inter-retrotransposon amplification polymorphism*) und iPBS (*inter-PBS amplification*) nutzen Insertionspolymorphismen transposabler Elemente und amplifizieren zu diesem Zweck Segmente zwischen zwei benachbarten Retrotransposons. Die verwendeten Primer hybridisieren im LTR (IRAP-Methode) oder in einer konservierten, von der Reversen Transkriptase für die Transposition benötigten Primerbindestelle (iPBS-Methode) der Retrotransposons und sind dabei in Richtung des Transposonendes orientiert.

Abbildung 12 zeigt exemplarisch den molekularen Fingerabdruck von 15 verschiedenen Kultur- und Wildreben, der mit einem der getesteten Primern angefertigt wurde. Insertionspolymorphismen des dem molekularen Fingerabdruck zugrundeliegenden *Copia*-Retrotransposons *Tvv1* resultieren in zahlreichen PCR-Produkten unterschiedlicher Größen, die bei der abgebildeten gelelektrophoretischen Auftrennung ein für jede Weinrebe einzigartiges Bandenmuster bilden.

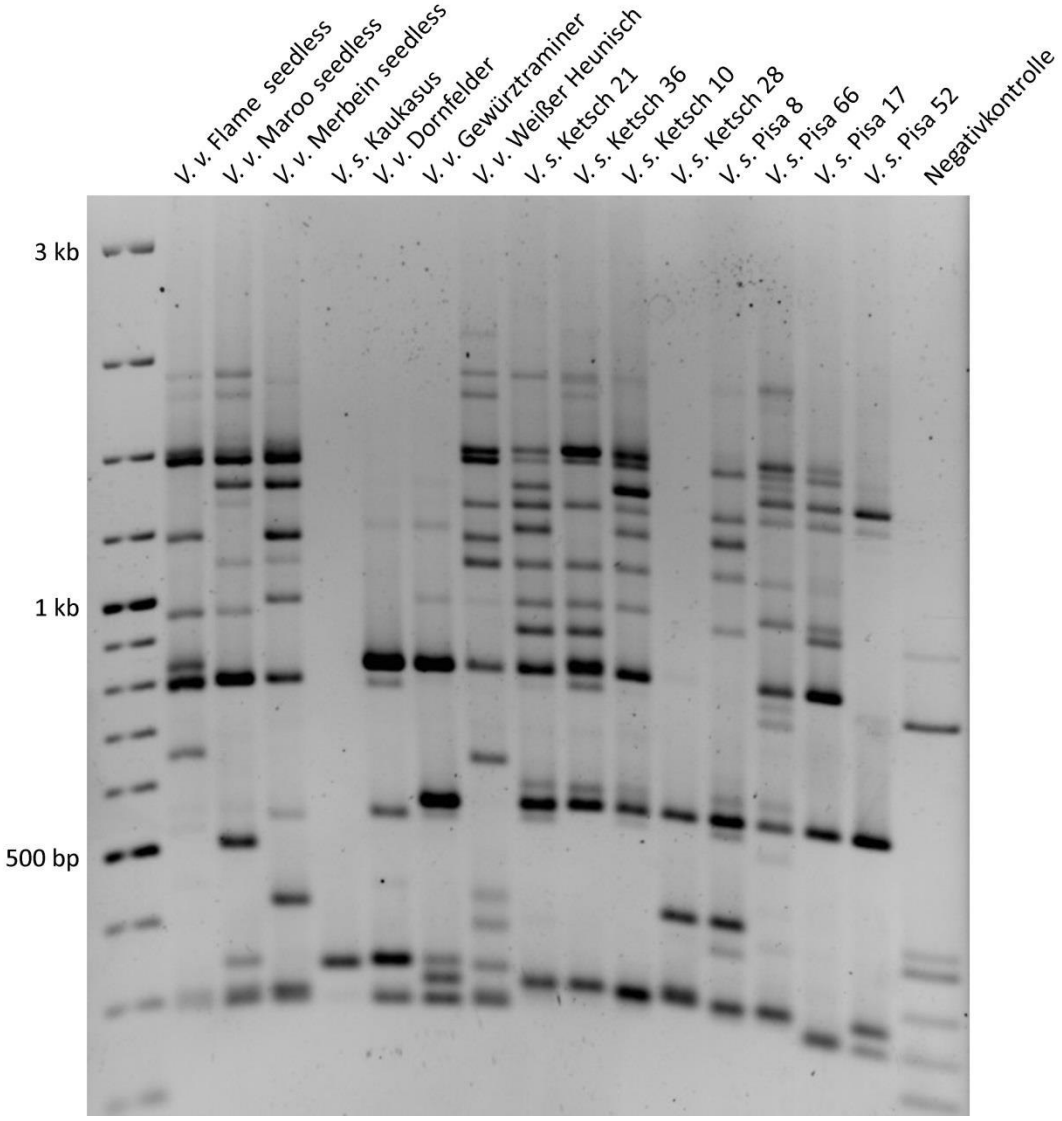

Abbildung 12: Molekularer Transposon-basierter Fingerabdruck verschiedener Kultur- und Wildreben

Exemplarisch dargestellt ist der molekulare Fingerabdruck von 15 verschiedenen Rebsorten, darunter drei Tafeltrauben (Flame seedless, Maroo seedless, Merbein seedless), drei für die Weinproduktion angebaute Edelreben (Dornfelder, Gewürztraminer, Weißer Heunisch) sowie neun Wilde Weinreben (Kaukasus, Ketsch 21, -36, -10, -28, Pisa 8, -66, -17, -53). Der verwendete Primer bindet im LTR des *Copia*-Retrotransposons *Tvv1* (D'Onofrio et al. 2010). Für alle Proben wurden Triplikate angefertigt und nur reproduzierbare Banden wurden in die weitere Auswertung übernommen. Banden, die auch in der Negativkontrolle auftraten, wurden verworfen.

Insgesamt wurden mit den drei in Tabelle 23 aufgeführten Primern 61 Banden erzeugt von denen 49 in den untersuchten Weinreben polymorph waren. Diese informativen Banden wurden zur Erstellung einer Binärmatrix verwendet, bei der „1" *Bande vorhanden* und „0" *Bande nicht vorhanden* entsprach.

Tabelle 23: Getestete Retrotransposon-basierte Marker

Typ	Primer	Banden total	polymorphe Banden
IRAP	*Tvv1*	22	22
IRAP	*Gret1*	14	5
iPBS	*F0100*	25	22
total		61	49

Auf Basis dieser Matrix wurde die in Abbildung 13 gezeigte Phylogenie der untersuchten Weinreben berechnet. Das Dendrogramm zeigt eine deutliche Unterteilung in die beiden Subspezies *Vitis vinifera* subsp. *vinifera* und *Vitis vinifera* subsp. *sylvestris*. Einzig die Wildrebe aus dem Kaukasus bildet eine Außengruppe. Innerhalb der Edelreben lassen sich zwei distinkte Cluster identifizieren. Beim ersten handelt es sich mit Dornfelder und Gewürztraminer um die Kulturreben, die für die Weinproduktion angebaut werden. Der zweite umfasst die drei Tafeltrauben Flame seedless, Maroo seedless und Merbein seedless. Bei den analysierten Wilden Weinreben aus Ketsch und Pisa ist keine genaue herkunftspezifische Einteilung möglich. Zwar bilden drei der vier Individuen aus Ketsch und zwei der vier Reben aus Pisa jeweils unabhängige Cluster, jedoch existiert darüber hinaus ein gemeinsamer Knotenpunkt der verbleibenden Indidivuen beider Populationen.

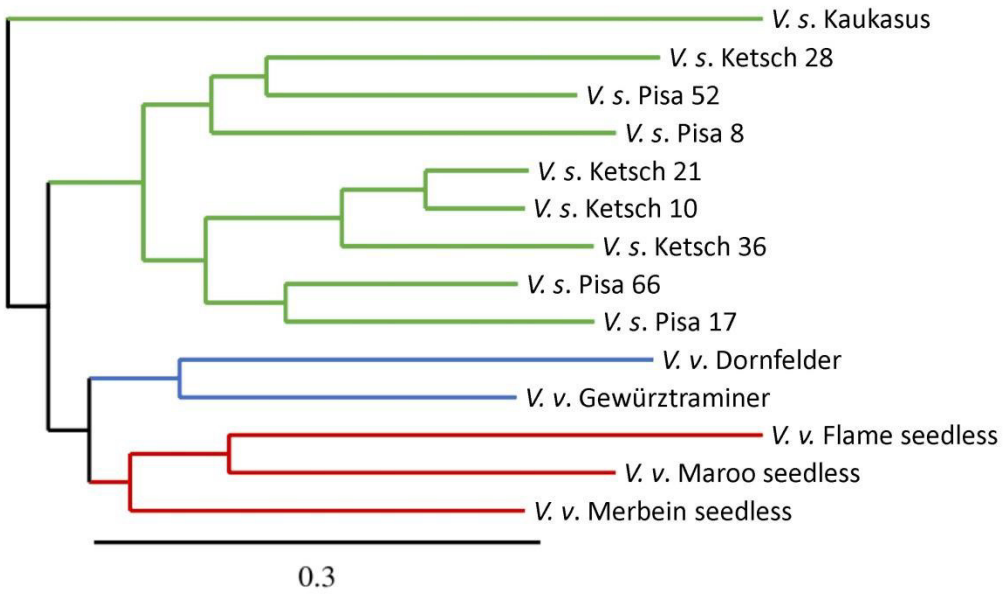

Abbildung 13: Transposon-basierte Phylogenie der untersuchten Weinreben
Das dargestellte Dendrogramm wurde mit dem Programm FAMD 1.31 mit dem Neighbor-Joining-Algorithmus berechnet. Die Kulturrebe Weißer Heunisch fehlt, da sich mit der iPBS-Methode keine Amplifikate erzeugen ließen. Astfarbe grün: *Vitis vinifera* subsp. *sylvestris*; blau: *Vitis vinifera* subsp. *vinifera* zur Weinproduktion; rot: Tafeltrauben der Spezies *Vitis vinifera* subsp. *vinifera*.

3.3 Identifizierung molekularer Fußspuren der Selektion im Genom der Wilden Weinrebe

3.3.1 Kartierung gegen das Referenzgenom

Die gefilterten Illumina-Sequenzdaten eines jeden Pools wurden gegen die aktuelle Version des Weinrebengenoms (12X.2) kartiert. Da das Referenzgenom auf Sequenzdaten eines Klons der Kulturrebe Pinot Noir basiert, wurde für die Kartierung der Wildrebendaten eine moderate Stringenz (*similarity* = 0,95; *length fraction* = 0,95) gewählt. *reads*, die unter diesen Voraussetzungen nicht im Referenzgenom positioniert werden konnten, wurden verworfen. Insgesamt wurden 48,96 Gb in Form von über 530 Mio. *reads* aligniert (Tabelle 24). Bei ca. 60 % dieser Sequenzen handelt es sich um sogenannte *paired reads*, also Sequenzpaare, deren *forward* und *reverse read* im korrekten Abstand zueinander positioniert werden können. Die Sequenzen decken je nach Pool zwischen 94,03 % und 95,45 % des 486 Mb langen Referenzgenoms ab. Die durchschnittliche Anzahl der *reads*, die eine Position im Genom abdecken, beträgt für die jeweiligen Pools zwischen 18,72 bis 33,92. Hierbei muss beachtet werden, dass bei der Kartierung jeweils beide Allele der einzelnen Individuen eines Pools abgedeckt werden sollen. Die Wilde Weinrebe ist ein diploider Organismus, dementsprechend muss die Abdeckung mindestens doppelt so groß sein wie die Anzahl der Individuen des Pools. Dies trifft auf alle Pools mit Ausnahme des Pools Pisa zu. Letzterer erreicht nur eine 18,72-fache Abdeckung und nicht die bei einer Individuenzahl von n=14 nötige minimale 28-fache Abdeckung des Genoms. Kombiniert man die Sequenzinformation aller Pools wird eine 104,52-fache Abdeckung des Rebengenoms erreicht.

Tabelle 24: Übersicht zur Kartierung gegen das Referenzgenom

Pool	n	*reads* gesamt	*reads* kartiert	% *paired reads*	%-uale Genom-abdeckung[1]	x-fache Genomab-deckung (*coverage*)[2]
Ketsch	14	225.795.243	154.451.467	60,68	94,19	31,00
Pisa	14	137.312.853	93.969.830	60,47	94,71	18,72
Frankreich	10	167.686.496	112.076.211	59,61	94,03	22,75
Spanien	15	251.036.846	171.313.872	60,59	95,45	33,92
total	53	781.831.438	531.811.380	60,39	96,35	104,52

[1] prozentualer Anteil des Referenzgenoms, der von mindestens einem *read* im Mapping abgedeckt wird; [2] durchschnittliche Anzahl der *reads*, die eine Position im Genom abdecken

Schlüsselt man die prozentuale Genomabdeckung chromosomenweise auf, zeigen sich nur geringfügige Unterschiede zwischen den einzelnen Chromosomen (Abbildung 14). Alle Chromosomen weisen eine Abdeckung von über 90 % auf, einzelne wie beispielsweise Chromosom 8 sind bis zu 98 % mit *reads* abgedeckt. Die einzige Ausnahme stellt das „Chromosom R“ dar, hier sind weniger als 85 % der Sequenz von *reads* der einzelnen Pools überspannt. Hierbei handelt es sich jedoch um kein reales Chromosom sondern um ein artifizielles Konstrukt aus allen *scaffolds* und *contigs*, die bis dato keinem der 19 Chromosomen der Weinrebe zugeordnet werden konnten (Canaguier et al. 2014).

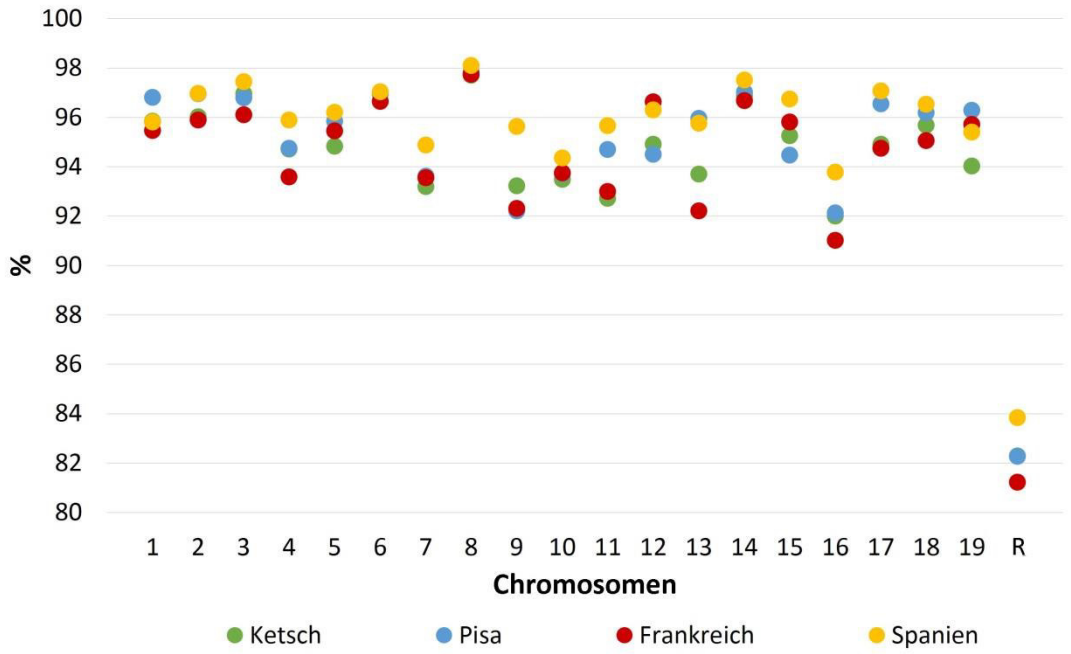

Abbildung 14: Abdeckung der einzelnen Chromosomen im *mapping*
Dargestellt ist der prozentuale Anteil eines Chromosoms, der von mindestens einem *read* in der Kartierung der verschiedenen Pools abgedeckt wird.

Weiterhin wurde getestet, ob die prozentuale Abdeckung eines Chromosoms mit dessen Gen- und GC-Gehalt korreliert. Hierzu wurde der Pearson-Korrelationskoeffizient r berechnet. Zwischen der prozentualen Abdeckung eines Chromosoms und dem Gehalt an Genen wurde eine starke positive Korrelation festgestellt ($r = 0{,}7844$; $p = 4{,}2 \cdot 10^{-5}$). Je höher der Gengehalt eines Chromosoms, desto besser ist es in der Kartierung durch *reads* abgedeckt. Als Berechnungsgrundlage für den Gengehalt diente der jeweilige prozentuale Anteil kodierender Sequenzen (*coding DNA sequence*, CDS) eines Chromosoms. Für den Zusammenhang zwischen der Abdeckung in der Kartierung und dem GC-Gehalt des

jeweiligen Chromosoms wurde ebenfalls eine positive Korrelation gemessen, diese fällt mit $r = 0{,}7321$ ($p = 2{,}4 \cdot 10^{-4}$) jedoch moderater aus.

Etwa 5 % des Genoms (22,1–29,0 Mb) werden bei der Kartierung von keinen Illumina-Sequenzen abgedeckt. Teils handelt es sich dabei um Bereiche, für die bereits im Referenzgenom keinerlei Sequenzinformation vorliegt. Da es sich nur um ein sogenanntes *draft genome* handelt, sind Lücken mit undefinierten Basen (N) nicht selten. Im Gesamtgenom beträgt der Anteil undefinierter Basen zwar nur 3,19 %, in den nicht-abgedeckten Bereichen der Kartierung ist der N-Anteil jedoch signifikant höher. Er beträgt je nach Pool zwischen 18,61 % und 36,13 %. Abgesehen davon weicht der GC-Gehalt dieser Bereiche nicht signifikant vom Restgenom ab, allerdings sind überdurchschnittlich viele repetitive Elemente und Sequenzen mit geringer Komplexität enthalten. Lücken, egal welcher Ursache sie sind, bereiten dem im weiteren Verlauf der Analyse verwendeten *sliding window* Probleme, da das Fenster fälschlicherweise Bereiche stromaufwärts und stromabwärts einer Lücke zusammenfasst. Um dies zu verhindern, müssen größere Lücken aus der Kartierung entfernt werden. Zu diesem Zweck wurde die Konsensussequenz des *mappings* extrahiert und an Positionen mit Lücken > 500 N in getrennte *contigs* zerlegt. Die resultierenden 5.737 Sequenzen dienten als Referenz für eine zweite Kartierung auf deren Basis die folgenden Analysen durchgeführt wurden. Sie weisen eine Gesamtlänge von knapp 471 Mb auf. Die 5.737 Referenzsequenzen befinden sich in Form einer Fasta-Datei im elektronischen Anhang.

3.3.2 Genomweite Identifizierung von Polymorphismen

Um die genetische Vielfalt innerhalb der vier Pools zu untersuchen, wurde auf Basis der kartierten Illumina-Sequenzdaten eine genomweite Suche nach Einzelnukleotid-polymorphismen (*single nucleotide polymorphisms*, SNPs) durchgeführt. Die Polymorphismen eines jeden Pools wurden mithilfe eines qualitätsbasierten Algorithmus identifiziert, der sowohl den Phred-Wert der untersuchten Position als auch die Güte der Basen in der unmittelbaren Umgebung berücksichtigt. Es wurden ausschließlich Bereiche betrachtet, die von mindestens 10 aber maximal 250 Illumina-Sequenzen abgedeckt sind. *reads*, die an mehr als einer Position in der Referenz kartieren, wurden von der Analyse ausgeschlossen. Um als SNP gewertet zu werden, musste eine Variante in wenigstens 5 % der Sequenzen der jeweiligen Position vorhanden sein. Alternativ genügte es, wenn 5 oder mehr *reads* die spezifische Variante unterstützten. Insgesamt wurden auf diese Art und

Weise in allen vier Pools kombiniert 23.417.203 SNPs identifiziert (Tabelle 25). Hiervon fielen ca. 9,6 Mio. SNPs auf den Pool Ketsch und 10,5 Mio. SNPs auf den Pool Pisa. Für die Weinreben aus Spanien und Frankreich wurden 13,4 Mio. bzw. 10,5 Mio. SNPs ermittelt. Daraus ergibt sich eine SNP-Frequenz von durchschnittlich einem SNP je 47,7 bp für den Pool Ketsch, 43,8 bp für Pisa und Frankreich sowie 34,6 bp für die Population aus Spanien. Die Listen aller identifizierten Polymorphismen inklusive der Anzahl und Sequenz der zugrundeliegenden Allele, deren Abdeckung in der Kartierung sowie ihre Allelfrequenzen befinden sich im elektronischen Anhang (Tabellen S1 – S4).

Tabelle 25: Zusammenfassung der identifizierten Polymorphismen

	Ketsch	Pisa	Frankreich	Spanien
Anzahl SNPs	9.629.132	10.540.327	10.464.656	13.415.960
1 SNP je ...	47,7 bp	43,8 bp	43,8 bp	34,6 bp
Transitionsmutationen	5.430.411	5.871.373	5.831.790	7.797.227
Transversionsmutationen	2.891.483	3.339.639	3.251.564	3.835.325
Ti/Tv	1,88	1,76	1,79	2,03
populationsspezifische SNPs	2.258.020	3.292.191	2.700.621	4.311.054
% populationsspezifische SNPs	23,45	31,23	25,81	32,13

Die identifizierten Polymorphismen verteilen sich auf alle 19 Chromosomen sowie das artifizielle Gerüstkonstrukt „Chromosom R“. In allen vier Pools zeigt Chromosom 18 die höchste Anzahl an SNPs, die wenigsten polymorphen Positionen beinhalten die Chromosomen 17, 11 und 2. Jedoch handelt es sich bei Chromosom 18 um das größte Chromosom der Weinrebe. Mit 34,6 Mb ist es 1,1- bis 1,8-mal so groß wie die verbleibenden Chromosomen. Bei den Chromosomen mit weniger Polymorphismen handelt es sich um die kleinsten Chromosomen.

Um eine bessere Vergleichbarkeit zwischen den Chromosomen zu gewährleisten, wurde daher für die einzelnen Chromosomen eines jeden Pools die SNP-Dichte berechnet (Abbildung 15). Hierfür wurde die Anzahl der SNPs eines Chromosoms durch die Länge der Konsensussequenz des entsprechenden Chromosoms dividiert. Es zeigt sich, dass die SNP-Dichte zwischen den Chromosomen eines Pools nur geringfügig schwankt und keine bemerkenswerten Unterschiede aufweist.

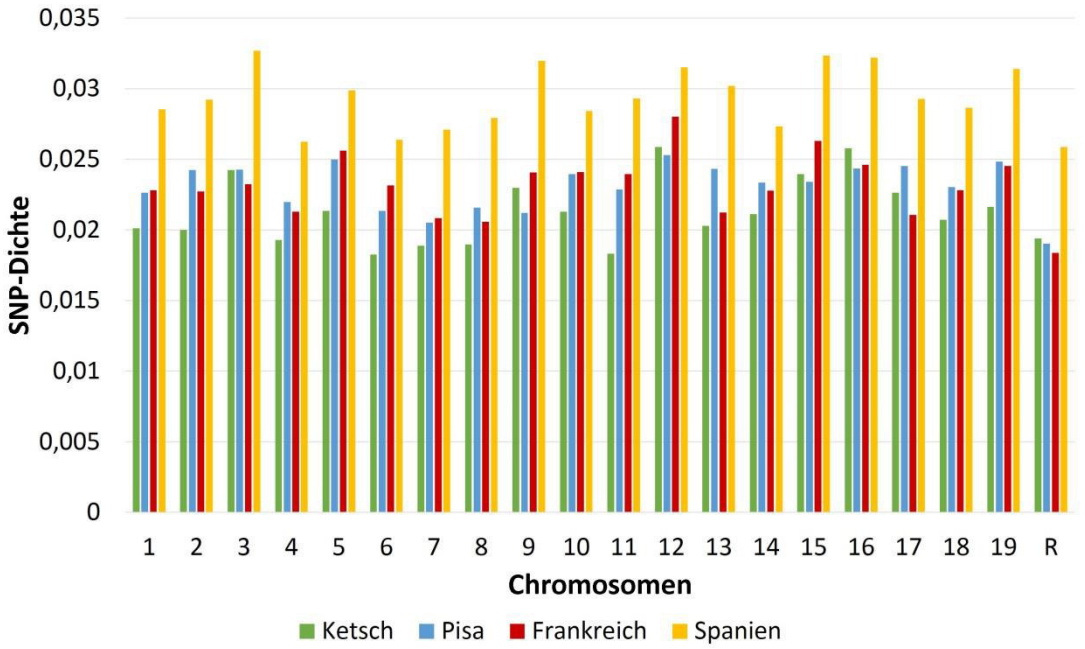

Abbildung 15: SNP-Dichte der einzelnen Chromosomen
Die im Diagramm abgebildete SNP-Dichte berechnet sich als Quotient aus der Zahl der SNPs, die auf einem Chromosom identifiziert wurden, und der Länge der Konsensussequenz des jeweiligen Chromosoms.

Es lassen sich verschiedene Formen von SNPs unterscheiden. Wird eine Purinbase gegen eine Purinbase oder eine Pyrimidinbase gegen eine Pyrimidinbase ausgetauscht, spricht man von einer Transitionsmutation. Bei Transversionsmutationen handelt es sich hingegen um einen Austausch einer Purinbase durch eine Pyrimidinbase oder umgekehrt. Neben diesen Substitutionen können einzelne Basen oder Gruppen von Basen deletiert oder inseriert vorliegen, diese Mutationsformen werden unter dem Begriff Indels zusammengefasst. Weiterhin wurden bei der Analyse komplexere Mutationen, die gekoppelt mehrere Nukleotide infolge betreffen, detektiert. Dazu zählen z. B. kleinere Inversionen. Der verwendete Algorithmus ist in der Lage, gekoppelte Polymorphismen bis zu einer Länge von 7 bp zu detektieren. Abbildung 16 zeigt beispielhaft anhand des Pools Ketsch die prozentuale Verteilung dieser verschieden Typen von Polymorphismen. Mit über 55 % bilden Transitionen den Großteil der identifizierten SNPs. Transversionen sind mit knapp 30 % deutlich seltener. Indels und komplexere Polymorphismen stellen mit unter 10 % bzw. 5 % den geringsten Anteil. Diese Verteilung ist in allen Pools nahezu identisch und weicht in den nicht dargestellten Pools nur durch wenige Prozentpunkte von dem hier gezeigten Beispiel des Pools Ketsch ab. Darüber hinaus wurde das Verhältnis von

Transitions- zu Transversionsmutationen (Ti/Tv) ermittelt. Es beträgt je nach Pool zwischen 1,76 und 2,03 (Tabelle 25).

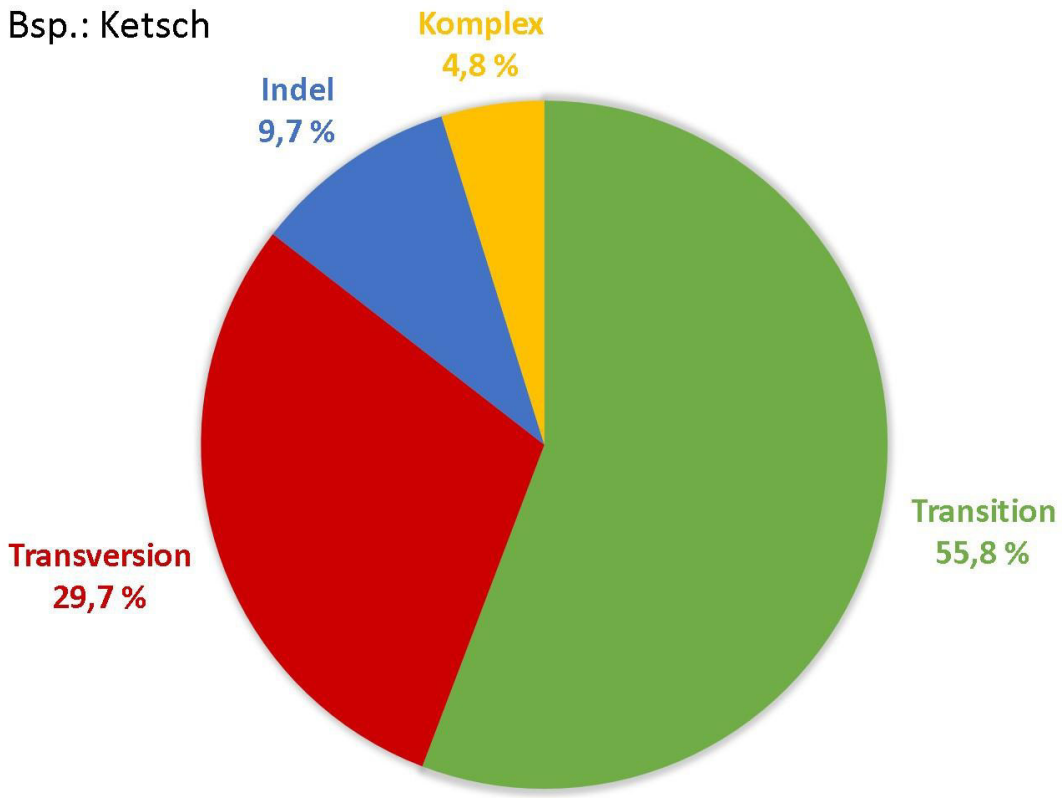

Abbildung 16: Verteilung der SNP-Typen
Das Tortendiagramm veranschaulicht exemplarisch für den Pool Ketsch die prozentuale Verteilung der verschiedenen Arten von Polymorphismen.

Anhand der identifizierten Polymorphismen wurde die Verteilung der minoren Allelfrequenzen (MAF) in den vier Populationen untersucht. Die MAF eines einzelnen SNPs beschreibt die Frequenz der seltensten Variante dieses Polymorphismus in der Gesamtpopulation. Berechnet wird sie bei der Hochdurchsatzsequenzierung von Pools daher als prozentualer Anteil der *reads*, die das seltenste Allel unterstützen, an der Gesamtheit der *reads*, die an dieser Position vorhanden sind. Abbildung 17 zeigt die Verteilung der minoren Allelfrequenzen im Pool Ketsch. Von den insgesamt 9,6 Mio. SNPs haben mehr als 3,7 Mio. SNPs eine MAF von ≤10 %. Der Großteil dieser SNPs weist eine MAF zwischen 5 und 10 % auf, nur ca. 30.000 Polymorphismen zeigen eine MAF von unter 5 %. Die verbleibenden fast 6 Mio. SNPs haben eine MAF von >10 %. Hierbei nimmt die Zahl der SNPs mit steigender MAF ab, minore Allelfrequenzen von 30–50 % sind weit weniger häufig als solche zwischen 10–30 %.

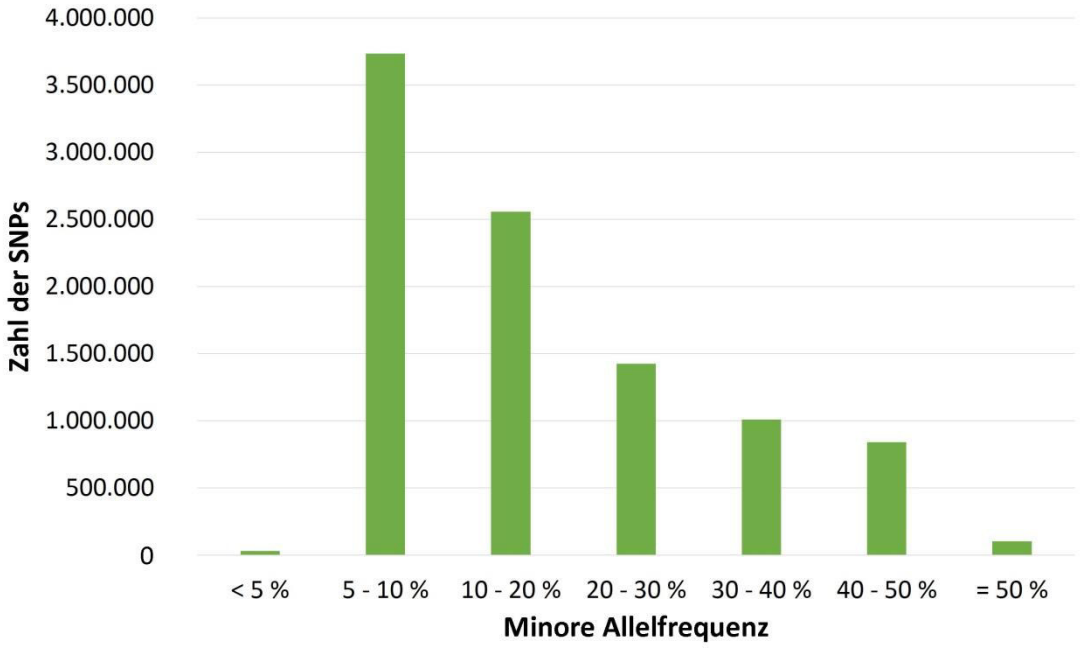

Abbildung 17: Verteilung der minoren Allelfrequenzen der identifizierten Polymorphismen
Die für den Pool Ketsch identifizierten SNPs werden gemäß ihrer minoren Allelfrequenz (MAF) auf sieben Kategorien verteilt. Die ersten beiden Kategorien umfassen einen Bereich von jeweils 5 %, das Intervall der folgenden vier Kategorien beträgt hingegen 10 %. Die verbleibende letzte Kategorie beinhaltet Polymorphismen mit identischer minoren und majoren Allelfrequenz.

SNPs, die nur in einem einzigen der vier untersuchten Pools identifiziert wurden, gelten als potenziell Pool-spezifische Polymorphismen. Positionen, die hingegen in mehreren oder sogar in allen vier Pools polymorph auftreten, werden als Pool-übergreifende Polymorphismen klassifiziert. Diese Relationen werden in Abbildung 18 in Form eines Venn-Diagramms veranschaulicht. Die Zahl Pool-spezifischer SNPs variiert in den einzelnen Pools von 2,3–4,3 Mio. SNPs. Den größten Anteil Pool-spezifischer SNPs weisen die Wilden Weinreben aus Spanien auf. 32,13 % aller im Pool Spanien identifizierten Polymorphismen treten ausschließlich in diesem Pool auf und finden sich in keiner der anderen untersuchten Populationen wieder. Dieser Anteil spezifischer SNPs ist im Pool Pisa mit 31,23 % ähnlich hoch. Die Wildrebenpopulationen aus Frankreich und Ketsch zeigen hingegen einen geringeren Prozentsatz Pool-spezifischer SNPs, hier machen die „einzigartigen“ Varianten nur 25,81 bzw. 23,45 % der Gesamtpolymorphismen aus.

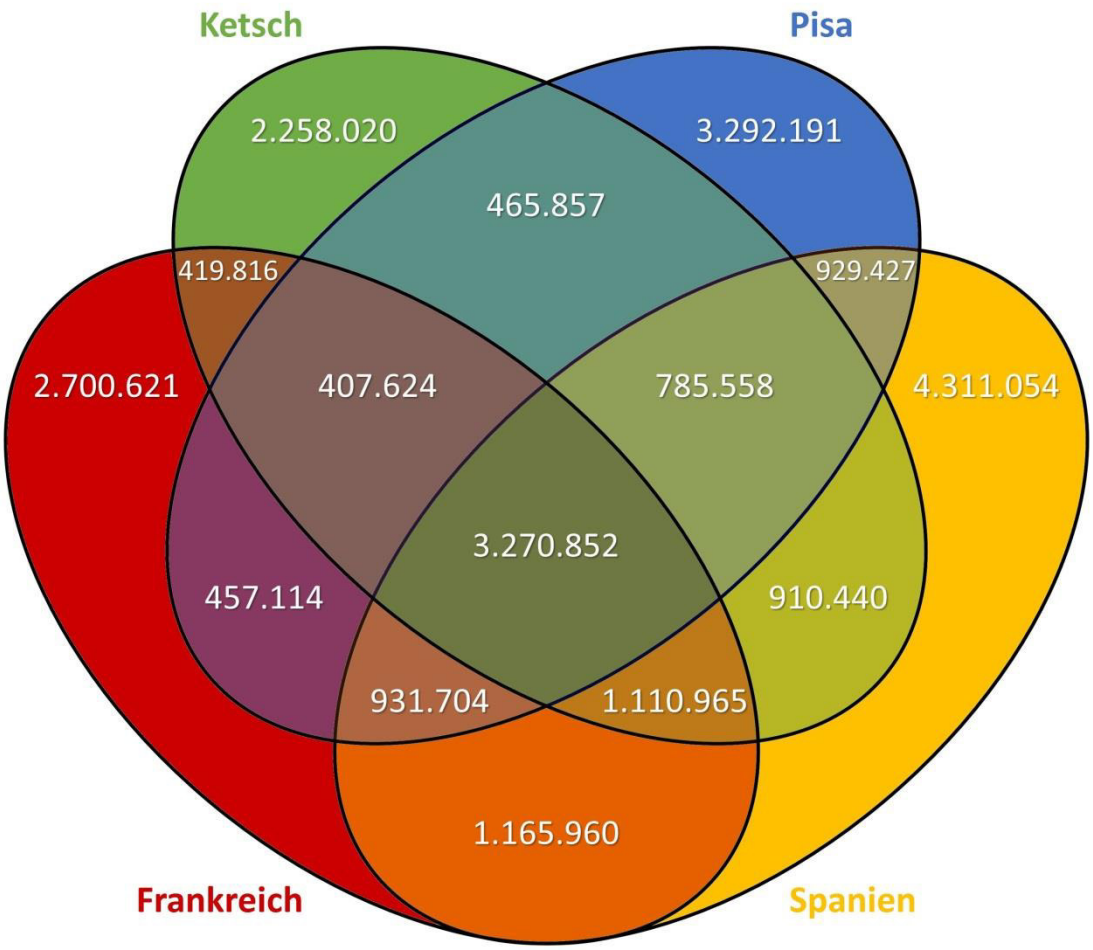

Abbildung 18: Venn-Diagramm Pool-spezifischer und Pool-übergreifender Polymorphismen
Das Diagramm zeigt die Verteilung der identifizierten Polymorphismen in den Pool-Sequenzierungen. Neben der Anzahl Pool-spezifischer Polymorphismen lässt sich die Anzahl jeder Kombination an Pool-übergreifenden Varianten ablesen.

Die größte Gemeinsamkeit in Form von Pool-übergreifenden SNPs weisen die Pools Spanien und Frankreich auf. An mehr als 6,4 Mio. Positionen sind die beiden Pools gleichermaßen polymorph. Die geringste Überlappung zeigen hingegen die Pools Ketsch und Pisa mit nur 4,9 Mio. gemeinsamen Polymorphismen. Die Schnittmenge der vier Pools bilden 3,3 Mio. SNPs, die in allen untersuchten Populationen vorhanden sind.

Da die Identifizierung der Polymorphismen ausschließlich *in silico* erfolgte, wurde eine Stichprobe von 15 zufällig ausgewählten SNPs experimentell validiert, um die Zuverlässigkeit der bioinformatischen Analysen abzuschätzen. Hierzu wurden die entsprechenden genomischen Abschnitte aller Individuen eines Pools mittels PCR amplifiziert und nach Sanger sequenziert. Um beide Extreme des Variationsspektrums abzudecken, wurden sowohl SNPs mit einer geringen minoren Allelfrequenz (MAF < 10 %) als auch Polymorphismen mit nahezu ausgeglichenen Allelfrequenzen (ca. 50 % / 50 %) überprüft. Alle 15 Polymorphismen konnten mittels Sanger-Sequenzierung bestätigt werden. Darüber hinaus wurden die auf Basis der Illumina-Daten bioinformatisch ermittelten Allelfrequenzen eines Pools mit den aus der Sanger-Sequenzierung abgeleiteten Genotypen der einzelnen Individuen des Pools verglichen. Es konnte eine signifikante Korrelation zwischen den Allelfrequenzen beider Methoden festgestellt werden (Pearson-Korrelationskoeffizient $r = 0{,}9918$; $p < 0{,}00001$). Die *in silico* berechneten Allelfrequenzen weichen im Durchschnitt nur um ±2,21 % von den experimentell bestätigten Frequenzen ab (Tabelle S5 im elektronischen Anhang).

3.3.3 Identifizierung von Kandidatenregionen unter Selektion

Es gibt zahlreiche Möglichkeiten, aktuelle oder vergangene Selektionsereignisse in einem Genom zu erkennen. Alle bedienen sich der Veränderungen in der Variationsstruktur der betroffenen genomischen Regionen. Hier soll die lokale Reduktion der durchschnittlichen Heterozygosität genutzt werden, um Kandidatenregionen unter Selektion im Genom der Wilden Weinrebe zu identifizieren. Grundlage für die Analysen bildet dabei die Berechnung der gepoolten Heterozygosität (H_P) nach Rubin et al. (2010). Die Werte werden mittels am Genom entlanggleitender Fenster erhoben. Für jedes Fenster wird die genetische Vielfalt des entsprechenden DNA-Abschnitts in Form des H_P-Wertes kalkuliert.

Zunächst musste die ideale Größe der Fenster ermittelt werden. Hierfür sollte die Verteilung der *contig*-Längen des Referenzgenoms berücksichtigt werden, da dieses bei der

vorangegangenen Kartierung der Illumina-Sequenzdaten bei Lücken > 500 N in getrennte *contigs* zerlegt wurde (vgl. Kapitel 3.3.1). H_P-Werte können nur für *contigs* berechnet werden, deren Länge mindestens der Fenstergröße entspricht. Kürzere *contigs* sind von den weiteren Analysen ausgeschlossen. Tabelle 26 veranschaulicht diesen Zusammenhang für sechs mögliche Fenstergrößen. Bei einer Fenstergröße von 10 kb fließen knapp 95 % des Referenzgenoms in die weitere Auswertung ein. Dieser Anteil verringert sich bei steigender Fenstergröße. Bei einer Fenstergröße von 40 kb können für ca. 90 % des Referenzgenoms H_P-Werte berechnet werden. Wählt man eine Fenstergröße von 60 kb, werden nur noch 88,11 % des Genoms analysiert.

Tabelle 26: Zusammenhang zwischen Fenstergröße und Länge der analysierten *contigs*

Fenster-größe	analysierte *contigs*	Gesamtlänge der analysierten *contigs*	% des Referenz-genoms
10 kb	>10 kb	460.770.155 bp	94,77
20 kb	>20 kb	454.731.774 bp	93,53
30 kb	>30 kb	447.475.511 bp	92,03
40 kb	>40 kb	441.419.952 bp	90,79
50 kb	>50 kb	435.393.922 bp	89,55
60 kb	>60 kb	428.392.649 bp	88,11

Ein weiterer Faktor, der bei der Ermittlung der optimalen Fenstergröße beachtet wurde, ist die durchschnittliche Anzahl der Polymorphismen, die sich in einem Fenster befinden. Wie in Abbildung 19 zu erkennen, besteht ein linearer Zusammenhang zwischen der Größe der Fenster und der Anzahl der SNPs je Fenster. Bei einer Fenstergröße von 10 kb befinden sich im Pool Ketsch im Schnitt 213 SNPs in jedem genetischen Abschnitt, der von einem Fenster analysiert wird. Vervierfacht man die Fenstergröße auf 40 kb so befinden sich durchschnittlich viermal so viele SNPs in einem Fenster. Bei 60 kb steigt die gemittelte SNP-Zahl je Fenster auf über 1.200. Bei einer Fenstergröße von 10 kb existieren vereinzelt Fenster, in denen keine Polymorphismen identifiziert wurden. Da die Allelfrequenzen der SNPs die Grundlage für die Berechnung der H_P-Werte nach Rubin et al. (2010) bilden, können für diese Fenster keine Werte kalkuliert werden. Unter Berücksichtigung von Tabelle 26 und Abbildung 19 wurde die Fenstergröße für die weiteren Analysen auf 40 kb festgelegt.

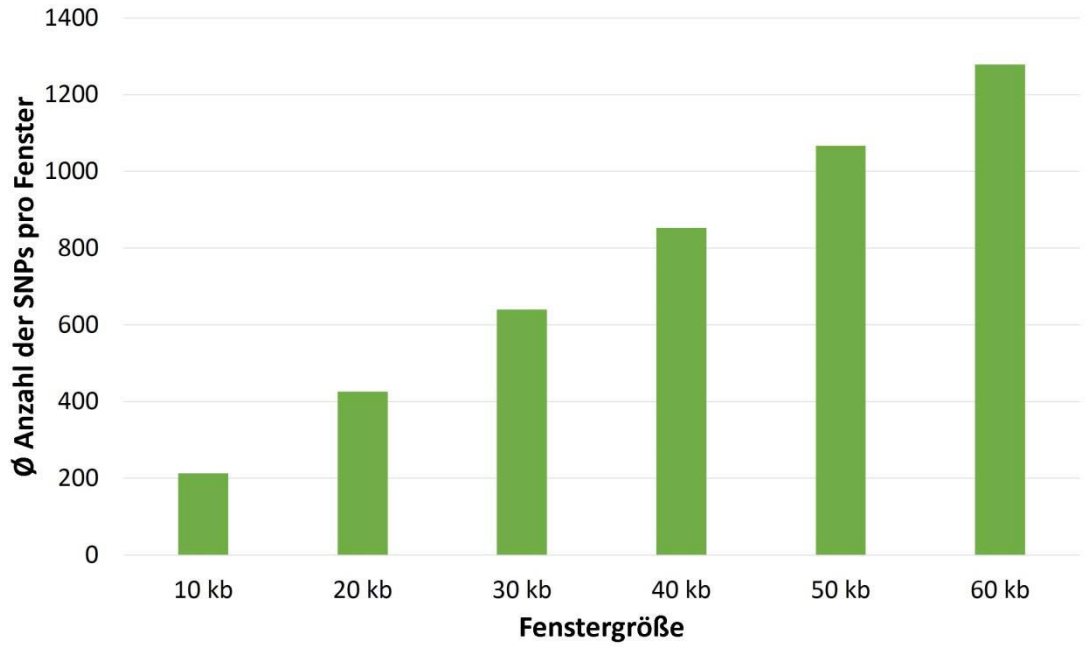

Abbildung 19: Korrelation zwischen der durchschnittlichen Anzahl der SNPs pro Fenster und der Fenstergröße
Das Genom wurde mit sechs möglichen Fenstergrößen gescannt und die Anzahl der SNPs pro Fenster ausgezählt. Im Balkendiagramm sind die Ergebnisse exemplarisch für den Pool Ketsch dargestellt.

Ein detaillierter Blick auf die SNP-Verteilung in den Fenstern am Beispiel des Pools Ketsch bei der gewählten Fenstergröße von 40 kb zeigt eine annähernde Normalverteilung um den Wert 854. Auffällig ist jedoch eine leichte Asymmetrie der Verteilung in Richtung höherer SNP-Zahlen je Fenster (Abbildung 20). Die Anzahl der SNPs pro Fenster reicht von 25 bis hin zu einer maximalen Anzahl von 3.632. Die verbleibenden drei Pools zeigen eine vergleichbare rechtsschiefe Normalverteilung um die jeweiligen Mittelwerte von 943 SNPs für Pisa und Frankreich sowie 1.186 SNPs pro Fenster für Spanien.

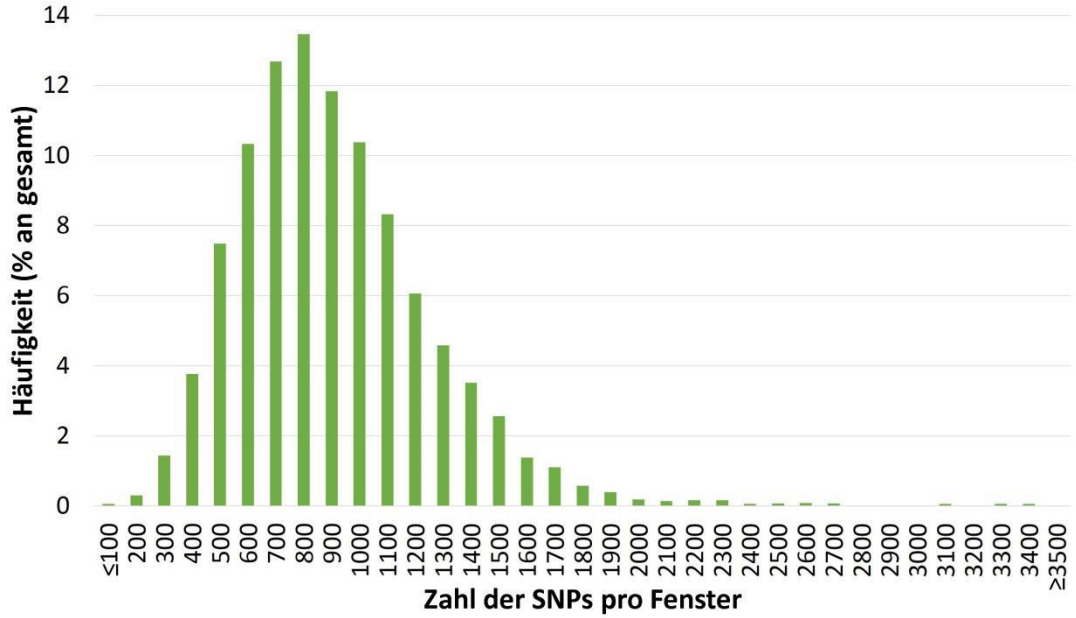

Abbildung 20: Zahl der SNPs pro Fenster für die gewählte Fenstergröße von 40 kb
Das Histogramm zeigt exemplarisch für den Pool Ketsch die Häufigkeitsverteilung der Zahl der SNPs pro Fenster bei der ausgewählten Fenstergröße von 40 kb.

Um eine ausreichend hohe Auflösung zu gewährleisten, muss das Fenster in überlappenden Schritten das Genom entlanggleiten. Daher wurde die Schrittgröße auf 10 kb festgelegt, was dafür sorgt, dass benachbarte Fenster zu jeweils 75 % überlappen.

Auf diese Art und Weise wurden für die Pools Ketsch und Pisa 36.429 bzw. 36.629 Fenster untersucht und entsprechend genauso viele korrespondierende H_P-Werte berechnet. Für die Pools Frankreich und Spanien betrug die Anzahl an Fenstern und somit an H_P-Werten 36.344 bzw. 36.943. Abbildung 21 zeigt für die einzelnen Pools die Häufigkeitsverteilung der berechneten H_P-Werte. In allen vier Fällen lässt sich eine glockenförmige Gestalt erkennen. Die Spanne, Mittelwerte (μ) und Standardabweichungen (σ) unterscheiden sich jedoch in den einzelnen Pools. Die größte Varianz von H_P-Werten weist hierbei der Pool Ketsch auf (0,1500-0,4361; μ=0,2964; σ=0,0474). Die H_P-Werte für den Pool Pisa erstrecken sich von 0,1519 bis 0,4124 (μ=0,2850; σ=0,0346). Der niedrigste H_P-Wert wurde für den Pool Frankreich gemessen, bei einer Spanne von 0,1484 bis 0,4234 (μ=0,2958; σ=0,0384). Den höchsten H_P- und Mittelwert zeigt der Pool Spanien (0,1743-0,4377; μ=0,3031). Zugleich weist dieser die kleinste Standardabweichung (σ=0,0342) auf.

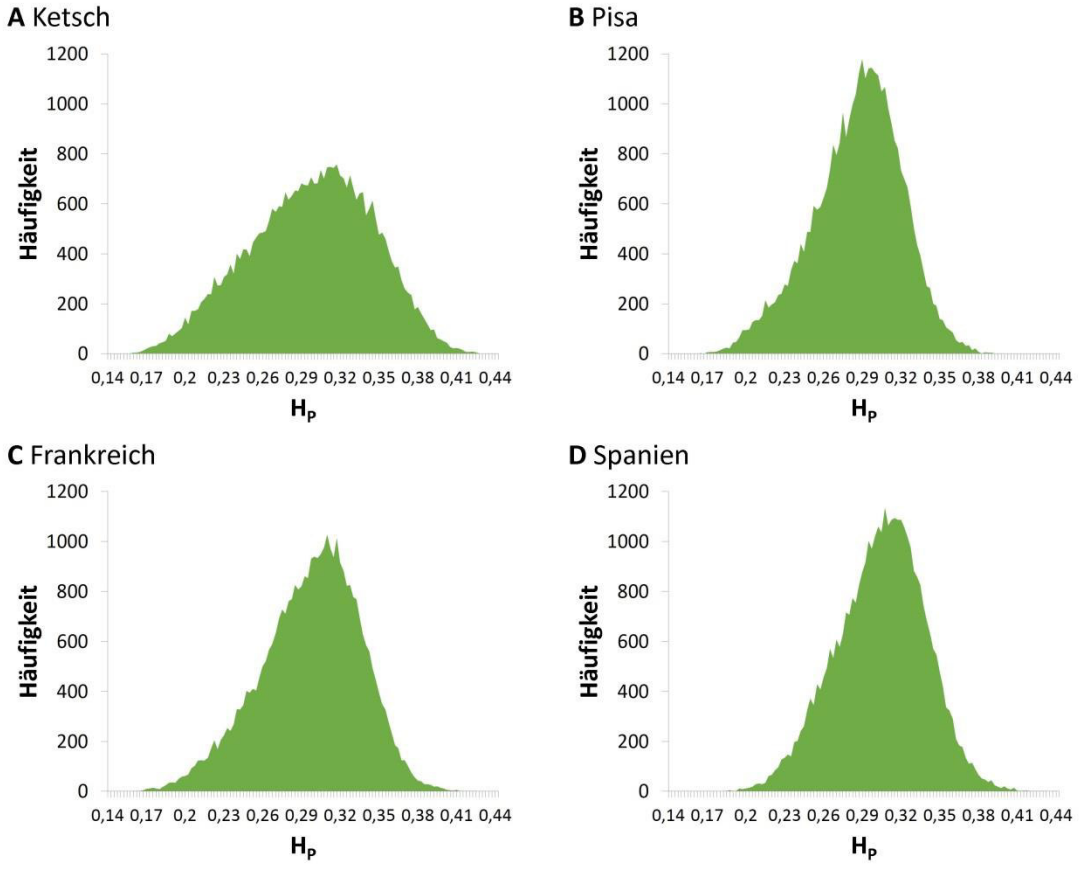

Abbildung 21: Häufigkeitsverteilung der H_P-Werte
Gezeigt sind die Histogramme der H_P-Werte bei Verwendung einer Fenstergröße von 40 kb und einer Schrittgröße von 10 kb für die Pools Ketsch (A), Pisa (B), Frankreich (C) und Spanien (D).

Zur besseren Visualisierung und Vergleichbarkeit zwischen den Pools wurden die H_P-Werte einer Z-Transformation unterzogen. Definitionsgemäß weisen die resultierenden ZH_P-Werte in allen Pools einen Mittelwert von 0 auf und besitzen eine Standardabweichung von 1 (Abbildung 22). Fenster mit reduzierter Heterozygosität erhalten demnach ein negatives, Fenster mit erhöhter Diversität ein positives Vorzeichen. Im Pool Ketsch bewegen sich die ZH_P-Werte zwischen -3,0854 und 2,9481, im Pool Pisa zwischen -3,8495 und 3,6867. Die ZH_P-Werte der Pools Frankreich und Spanien erstrecken sich von -3,8379 bis 3,3254 bzw. von -3,7642 bis 3,9319. Die ZH_P-Werte folgen ähnlich wie die H_P-Werte annähernd einer Normalverteilung. Alle berechneten H_P- und ZH_P-Werte befinden sich in tabellarischer Form im elektronischen Anhang (Tabellen S6 – S9).

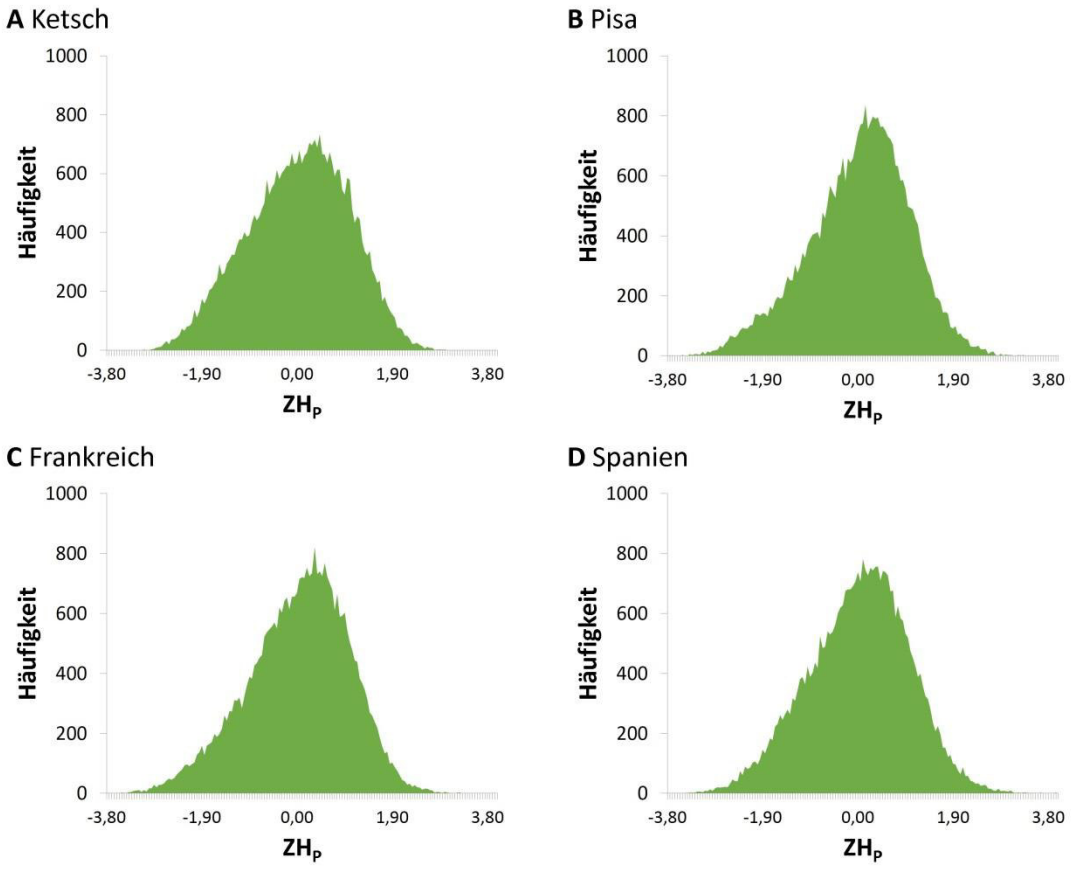

Abbildung 22: Häufigkeitsverteilung der ZH_P-Werte
Dargestellt sind die Histogramme der durch die Z-Transformation erhaltenen ZH_P-Werte bei Verwendung einer Fenstergröße von 40 kb und einer Schrittgröße von 10 kb für die Pools Ketsch (A), Pisa (B), Frankreich (C) und Spanien (D).

In Abbildung 23 sind die ermittelten ZH_P-Werte für die einzelnen Pools entlang der Chromosomen aufgetragen. Jeder Datenpunkt im Diagramm entspricht einem analysierten Fenster mit dem korrespondierenden ZH_p-Wert auf der Y-Achse. Zur besseren Differenzierung sind die einzelnen Chromosomen farbig voneinander abgehoben. Es ist zu erkennen, dass die genetische Diversität in Form der ZH_p-Werte entlang der einzelnen Chromosomen stark schwankt. Es existieren genetische Abschnitte mit erhöhten und Regionen mit erniedrigten ZH_p-Werten. Die Signale zeigen hierbei einen charakteristischen Verlauf: Die Veränderung der ZH_p-Werte erfolgt nicht schlagartig von einem Fenster zum nächsten. Die Zu- bzw. Abnahme geschieht vielmehr graduell über mehrere Fenster hinweg.

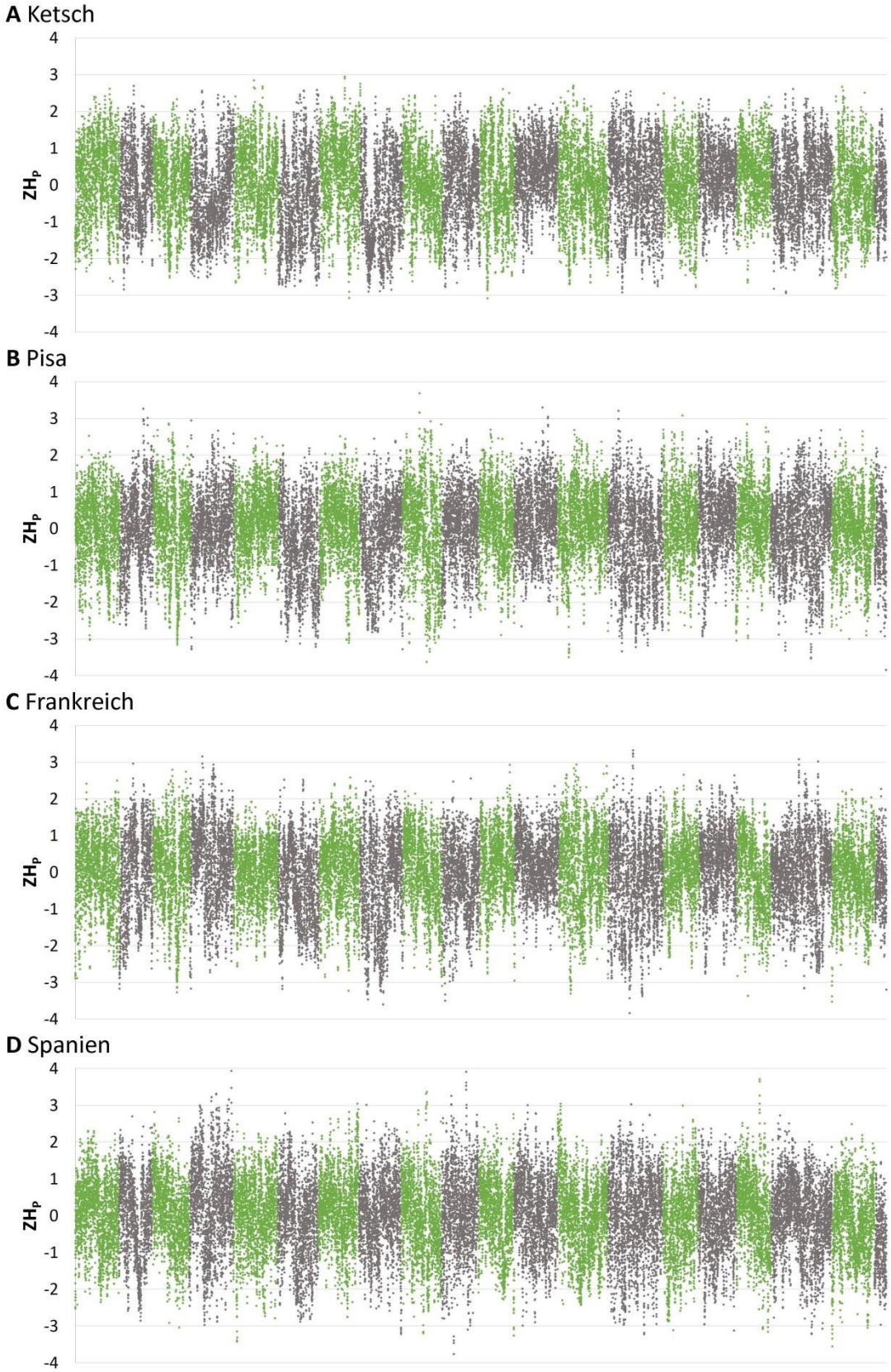

Abbildung 23: Genomweite Verteilung der ZH_P-Werte in den verschiedenen Pools
Die genomische Position ist auf der X-Achse gegeben. Die einzelnen Chromosomen sind farbig dargestellt. Jeder Datenpunkt entspricht einem Fenster mit dem zugehörigen ZH_P-Wert auf der Y-Achse.

Da *selective sweeps* mit einer starken Reduktion der genetischen Diversität einhergehen, stellen die Fenster, die die Minima der ZH_P-Verteilung bilden, den Ausgangspunkt für die weiteren Analysen dar. Hierfür wurde zunächst für jeden Pool ein geeigneter Schwellenwert festgelegt, bei dessen Unterschreitung ein Fenster als signifikant und der zugrundeliegende genetische Abschnitt als Kandidatenregion für ein Selektionsereignis gewertet wird. Mithilfe eines statistischen Tests wurde daher überprüft, ob solche erniedrigten H_p-Werte zufällig, das heißt in Abwesenheit von Selektion, entstehen könnten. Zu diesem Zweck wurde der Zufall durch eine Permutationsmethode simuliert: Die genomische Struktur in Form der SNP-Positionen wurde beibehalten, die im Datensatz erhobenen Allelfrequenzen wurden jedoch gemischt und zufällig auf die Positionen verteilt (*shuffle*). Auf Basis dieses permutierten Datensatzes wurden im Anschluss die H_p-Werte berechnet und der niedrigste H_p-Wert gespeichert. Diese Simulation wurde für jeden Pool 10.000-fach wiederholt und dementsprechend die 10.000 durch Zufall entstandenen, niedrigsten H_p-Werte aufgezeichnet. In Abbildung 24 sind diese 10.000 niedrigsten H_p-Werte im Vergleich zu den realen H_p-Werten des Pools Ketsch aufgetragen. Anhand dieser Verteilung lässt sich ein gewünschter Schwellenwert ablesen. Als Schwellenwert wurde der zehntkleinste Wert der dargestellten Verteilung gewählt. Anders ausgedrückt, ein solcher Wert kommt im entsprechenden Datensatz nur einmal in 1.000 durch Zufall zustande, was einem Signifikanzniveau von $p \leq 0{,}001$ entspricht. Der niedrigste Schwellenwert wurde für den Pool Ketsch mit $H_p=0{,}1911$ festgelegt, gefolgt von Pool Pisa mit $H_p=0{,}2043$. Die Schwellenwerte der Pools Frankreich und Spanien betragen 0,2115 bzw. 0,2396.

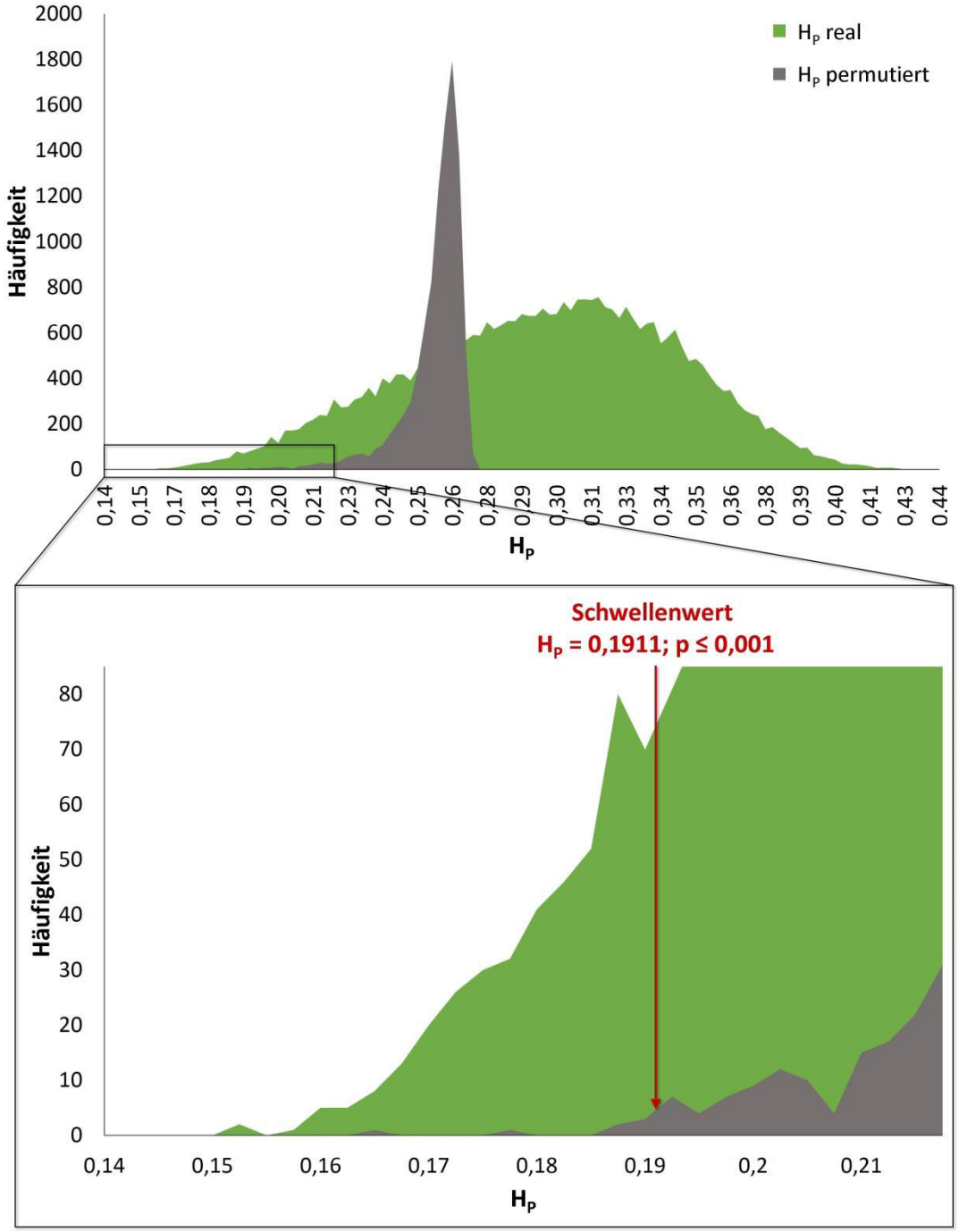

Abbildung 24: Ermittlung des Schwellenwertes für ein Selektionsereignis
Dargestellt sind exemplarisch für den Pool Ketsch die Verteilung der H_P-Werte des realen Datensatzes (H_P real, grün) und die während der Simulation aufgezeichneten 10.000 niedrigsten H_P-Werte (H_P permutiert, grau). Der vergrößerte Ausschnitt zeigt die Position des Schwellenwertes bei einem Signifikanzniveau von p≤0,001. Es handelt sich um den zehntniedrigsten Wert der permutierten Daten.

Abbildung 25 liefert einen genomweiten Überblick zur Verteilung der Fenster mit negativen ZH_P-Werten. Fenster, die den festgelegten Schwellenwert des jeweiligen Pools unterschreiten, sind rot markiert. Insgesamt weisen 3.373 Fenster einen signifikant reduzierten H_P- und entsprechenden ZH_P-Wert auf. Hiervon stammen 466 Fenster aus dem Pool Ketsch und 653 Fenster aus dem Pool Pisa. Im Pool Frankreich unterschreiten 838 Fenster den ermittelten Schwellenwert. Mit 1.416 Fenstern weist der Pool Spanien die größte Zahl genetischer Abschnitte mit signifikant reduzierter Diversität auf. Die 3.373 signifikanten Fenster verteilen sich über alle Chromosomen der Pools. Eine Ausnahme stellt das Chromosom 16 dar, hier befinden sich beim Pool Ketsch keine Fenster, die den Schwellenwert unterschreiten. Neben dem angesprochenen Chromosom 16 weisen die Chromosomen 12 und 7 in allen vier Pools die geringste Anzahl signifikanter Fenster auf. Die größte Zahl Fenster mit signifikant erniedrigter Diversität findet sich auf den Chromosomen 14, 8 und 6. Es lässt sich keine signifikante Korrelation zwischen der Anzahl der identifizierten Fenster mit reduzierter Diversität und der Chromosomenlänge feststellen.

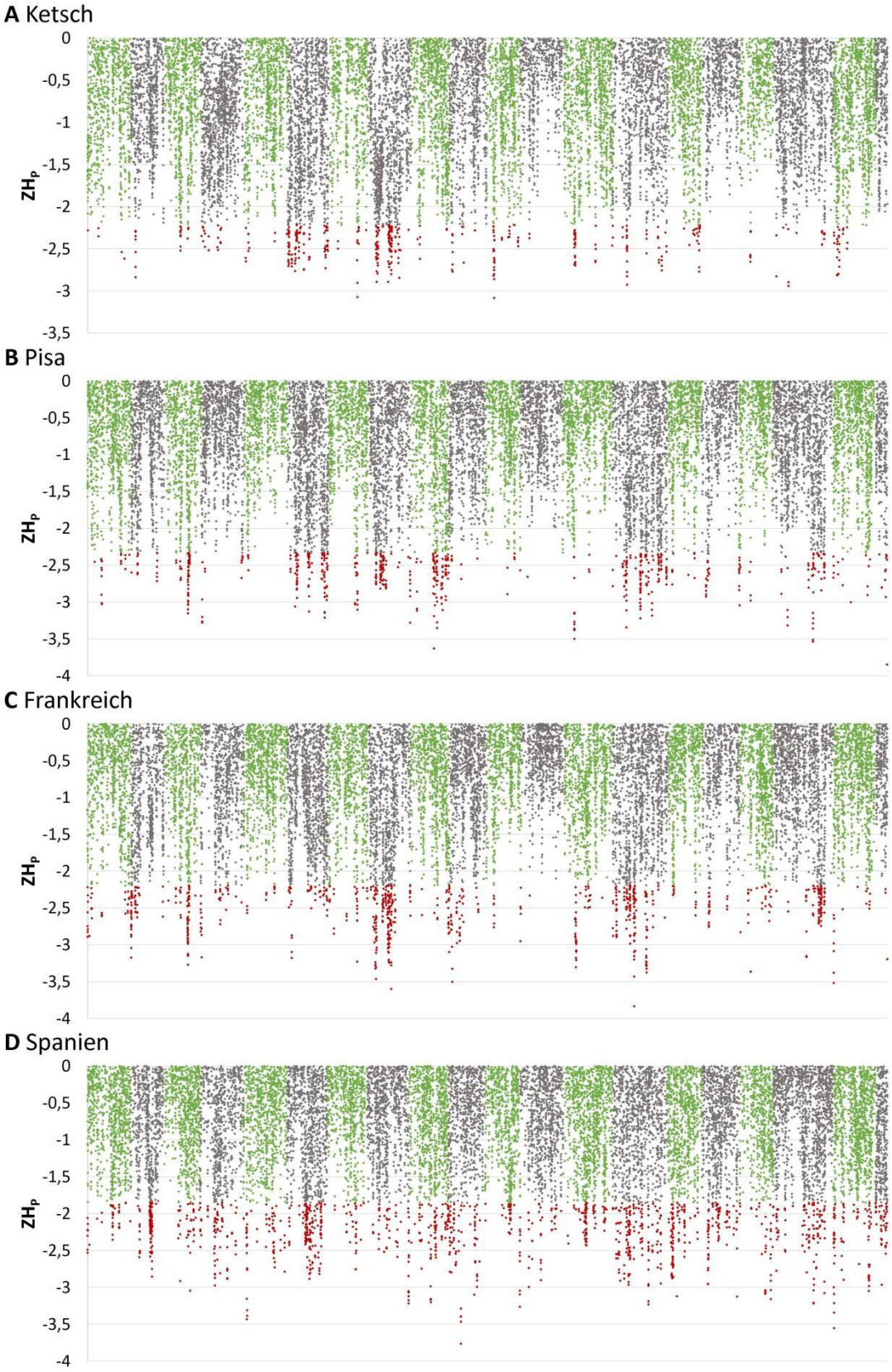

Abbildung 25: Identifizierung der Kandidatenregionen mit signifikant reduzierter Diversität
Dargestellt ist der negative Abschnitt der genomweiten Verteilung der ZH_P-Werte. Fenster, die den festgelegten Schwellenwert des jeweiligen Pools unterschreiten, sind rot markiert.

Unterschreiten unmittelbar benachbarte Fenster gleichermaßen den festgelegten Schwellenwert, werden sie zu einem gemeinsamen Signal zusammengefasst. Auf diese Art und Weise ergeben sich für den Pool Ketsch 153 und für den Pool Pisa 198 Signale. In den Pools Frankreich und Spanien lassen sich die identifizierten Fenster zu 239 bzw. 434 Signalen vereinen. Mithilfe einer BLAST-Suche wurden die den Signalen zugrundeliegenden Kandidatenregionen im publizierten Referenzgenom des Pinot Noir Klons PN40024 identifiziert (Tabelle 27). Die Signale weisen eine durchschnittliche Länge von 63,89 kb (Ketsch) bis 67,97 kb (Spanien) auf. Die Gesamtlänge aller Signale beträgt im Pool Ketsch 9,77 Mb und im Pool Pisa 13,09 Mb, was einem prozentualen Anteil am Genom von 2,01 bzw. 2,69 % entspricht. Für den Pool Frankreich ergibt sich eine Gesamtlänge aller Signale von 16,09 Mb, was mit 3,31 % des Genoms korrespondiert. Den größten prozentualen Anteil am Genom erreichen mit 6,07 % die Signale des Pools Spanien. Ihre Gesamtlänge beträgt 29,50 Mb. Tabelle S10 im elektronischen Anhang fasst detailliertere Positions- und Längeninformationen zu jedem Signal der einzelnen Pools zusammen. Darüber hinaus liegen die Sequenzen der Signale in Form einer Fasta-Datei vor.

Tabelle 27: Übersicht der identifizierten Kandidatenregionen

Pool	Schwellen-wert	Anzahl Fenster	Anzahl Signale	Ø Länge der Signale (bp)	Gesamtlänge aller Signale (bp)	Anteil am Genom (%)
Ketsch	0,1911	466	153	63.886,69	9.774.663	2,01
Pisa	0,2043	653	198	66.135,48	13.094.826	2,69
Frankreich	0,2115	838	239	67.332,29	16.092.417	3,31
Spanien	0,2396	1.416	434	67.974,16	29.500.787	6,07

3.3.4 Funktionelle Annotation der Kandidatenregionen

In den durch Analyse der gepoolten Heterozygosität identifizierten Kandidatenregionen der vier Pools sind zusammengenommen 2.601 verschiedene proteinkodierende Gene annotiert (V0 Annotation des 12X.0 Referenzgenoms). Im Pool Ketsch sind in 148 Signalen insgesamt 799 Gene enthalten. 5 Kandidatenregionen (3,27 %) weisen hingegen keine Gene auf. Der Anteil dieser genlosen Positionen ist im Pool Pisa mit 14,14 % deutlich höher. Hier befinden sich in nur 170 der 198 analysierten Signale ein oder mehrere Gene. Zusammenfassend lassen sich im Pool Pisa 704 Gene identifizieren. Für den Pool Frankreich wurden in 218 der 239 erfassten Loci 929 verschiedene Gene lokalisiert. Die übrigen 21 Signale (8,79 %) weisen keine Annotation auf. Die größte Anzahl proteinkodierender Gene

wurde im Pool Spanien identifiziert. Die 1.159 Gene verteilen sich auf 323 der 434 untersuchten Kandidatenregionen. In den verbleibenden 111 Signalen (25,58 %) sind jedoch keine Gene annotiert. Im Vergleich mit den anderen drei Pools handelt es sich hierbei um den höchsten Anteil genloser Kandidatenregionen. Eine Auflistung aller in den einzelnen Pools identifizierten Gene befindet sich in Form der Tabelle S11 im elektronischen Anhang.

Auf Basis der Anzahl der annotierten proteinkodierenden Gene und der Gesamtlänge der zugrundeliegenden genetischen Loci wurde der Gengehalt oder die Gendichte der analysierten Kandidatenregionen berechnet. Für den Pool Ketsch ergibt sich so eine Dichte von einem Gen je 12,2 kb. Der Gengehalt der Signale des Pools Ketsch ist somit im Vergleich zum Gesamtgenom, in dem im Durchschnitt je 18,6 kb ein Gen annotiert ist, erhöht. Die Kandidatenregionen der Pools Frankreich und Pisa weisen hingegen mit einem Gen je 17,3 bzw. 18,6 kb eine dem Gesamtgenom ähnliche Gendichte auf. Den geringsten durchschnittlichen Gengehalt findet man in den Signalen des Pools Spanien. Hier ist im Schnitt nur alle 25,5 kb ein Gen annotiert.

Mithilfe des in Abbildung 26 gezeigten Venn-Diagramms wurde untersucht, ob es bei den identifizierten Kandidatengenen Überlappungen zwischen den einzelnen Pools gibt. Analog zur Analyse der Polymorphismen (Abbildung 18, Kapitel 3.3.2) gelten hierbei Gene, die nur in einem einzigen der vier untersuchten Pools identifiziert wurden, als Pool-spezifische Kandidatengene. Proteinkodierende Gene, die hingegen in mehreren oder sogar in allen vier Pools erfasst wurden, werden als Pool-übergreifende Kandidatengene bezeichnet. Die Zahl Pool-spezifischer Kandidatengene variiert in den einzelnen Pools von 268 bis 769 Genen. Den größten Anteil Pool-spezifischer Kandidatengene weisen die Wilden Weinreben aus Spanien auf. 66,35 % aller im Pool Spanien identifizierten Gene tauchen ausschließlich in diesem Pool auf und finden sich in keiner der anderen untersuchten Populationen. Mit 56,57 % sind auch im Pool Ketsch mehr als die Hälfte der identifizierten Kandidatengene Pool-spezifisch. Die Wildrebenpopulationen aus Pisa und Frankreich zeigen hingegen einen geringeren Prozentsatz Pool-spezifischer Kandidatengene, hier macht diese Gruppe von Genen nur 38,12 bzw. 38,75 % der Gesamtheit der im jeweiligen Pool identifizierten Gene aus. Von den insgesamt 2.601 putativ selektierten Kandidatengenen sind 752 Gene in mehr als einem Pool vorhanden. Den größten Überlapp zeigen genau wie bei den Pool-übergreifenden Polymorphismen die Wilden Weinreben aus Spanien und Frankreich. Die

geringsten Gemeinsamkeiten zeigen hingegen die Pools Ketsch und Spanien, gefolgt von Ketsch und Pisa. Die Schnittmenge aller vier Pools bilden 33 Kandidatengene, die in allen untersuchten Populationen vorhanden sind und in den folgenden Ausführungen daher als „Kerngene“ (*core genes*) bezeichnet werden.

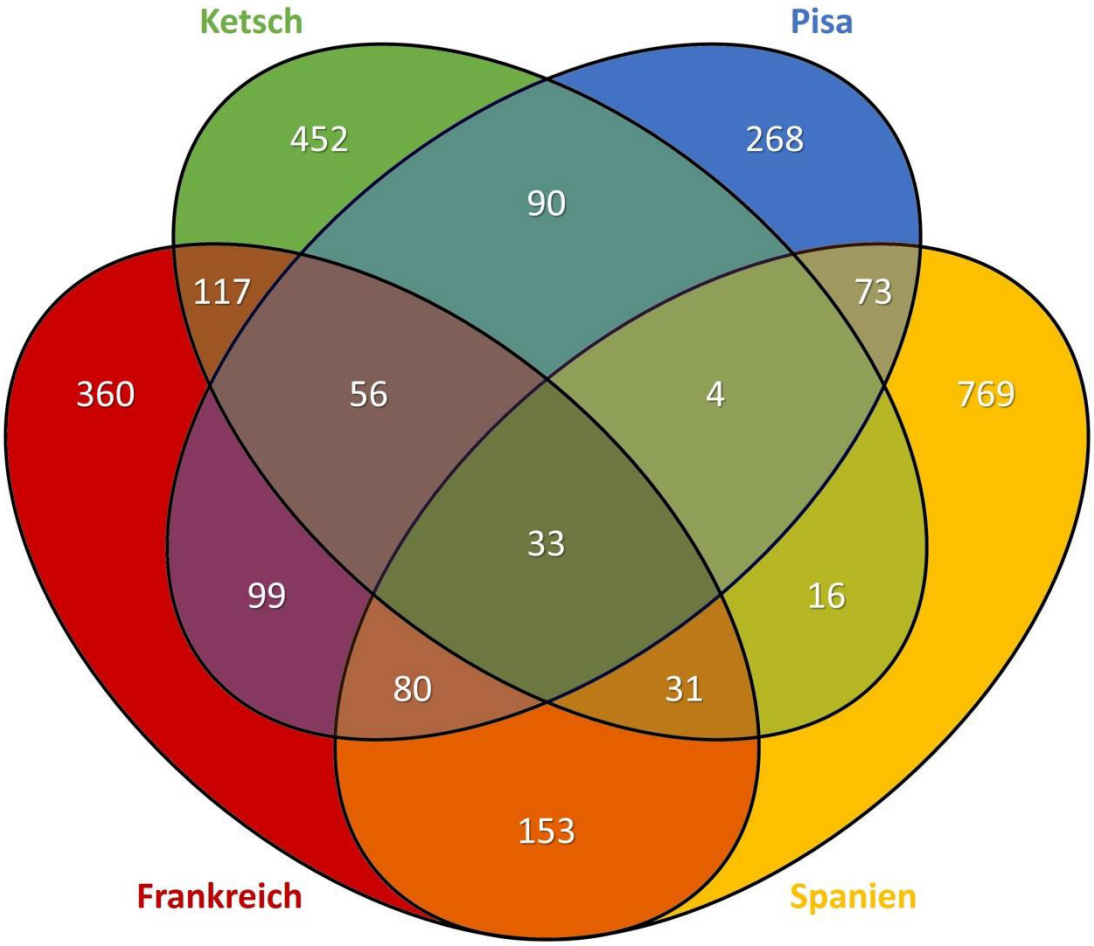

Abbildung 26: Venn-Diagramm Pool-spezifischer und Pool-übergreifender Kandidatengene
Das Diagramm zeigt die Verteilung aller identifizierter Kandidatengene. Neben der Anzahl Pool-spezifischer Gene lässt sich die Anzahl jeder Kombination an Pool-übergreifenden Genen ablesen. 33 Kandidatengene sind in allen vier Pools enthalten und werden daher als „Kerngene“ bezeichnet.

Die weiterführende Charakterisierung der identifizierten Kandidatengene erfolgte durch eine GO-Annotation mit dem Online-Tool PLAZA. Hierbei wurden die Gene funktionellen Gruppen (*Gene Ontology-Terms*) zugeordnet, wodurch Rückschlüsse auf ihre molekulare Funktion und ihre Beteiligung an biologischen Prozessen möglich sind. Insgesamt konnten 2.020 der 2.601 Kandidatengene mit *GO-Terms* annotiert werden. Angesichts der Tatsache, dass ein Gen mehreren Ontologien zugeordnet werden kann, ergibt sich für die 2.020 analysierten Kandidatengene eine Gesamtzahl von 3.461 *GO-Terms*. Bei den verbleibenden

581 Genen ohne GO-Zuordnung handelt es sich zumeist um Gene mit unbekanntem Proteinprodukt, für die auch in anderen Pflanzenspezies keine Orthologen identifiziert werden konnten.

Aufgrund der großen Anzahl von 2.020 Kandidatengenen mit 3.461 *GO-Terms* erschien eine detailliertere manuelle Auswertung dieser Gene nicht sinnvoll. Daher wurden im Weiteren vier Auswertestrategien verfolgt, um die Zahl der analysierten Gene einzugrenzen. Zunächst lag der Fokus auf den fünf Kandidatenregionen mit den niedrigsten ZH_P-Werten der einzelnen Pools, gefolgt von den 33 Kerngenen, die in allen vier Populationen identifiziert wurden. Ferner wurden statistisch angereicherte *GO-Terms* sowie Gengruppen mit besonderer Relevanz für die Züchtung betrachtet.

Analyse der Gene in den Top 5 Kandidatenregionen

Als Top 5 Kandidatenregionen wurden die fünf Fenster eines jeden Pools betrachtet, die in der vorangegangenen genomweiten Analyse der gepoolten Heterozygosität die extremen Minima der ZH_P-Verteilung bildeten. Im Gegensatz zur Gesamtheit der Signale, bei denen bis zu einem Viertel der Kandidatenregionen keine Gene beinhalten, ist in den Top 5 Fenstern in allen vier Pools mindestens ein Gen annotiert. Im Durchschnitt sind es 3,1 Gene. Den höchsten Gengehalt zeigt das viertplatzierte Fenster des Pools Frankreich mit sieben Genen. Tabelle 28 listet die Gene in der Rangfolge ihrer zugrundeliegenden Kandidatenregionen poolweise auf. Neben ZH_P-Wert, Position im Genom und Gen-ID ist, soweit bekannt, auch das resultierende Proteinprodukt angegeben. Unter den insgesamt 62 Genen befinden sich wichtige Haushalts- und Strukturgene wie Gene für ribosomale Proteine oder Bestandteile des Lichtsammelkomplexes der Chloroplasten. Andere lassen eine Zugehörigkeit zu Transportprozessen vermuten oder stehen im Zusammenhang mit der Regulation der Genaktivität. Insgesamt sind in Tabelle 28 zwar 62 Gene aufgeführt, jedoch sind davon einige doppelt in der Liste. So tauchen beispielsweise die drei Gene GSVIVG01021482001 (*squalene-hopene cyclase*), GSVIVG01021483001 (*beta-amyrin synthase*) und GSVIVG01021484001 (*beta-amyrin synthase*) sowohl im erst- als auch im drittplatzierten Fenster des Pools Spanien auf. Darüber hinaus sind die Gene des im Pool Frankreich drittplatzieren Fensters mit denen des zweitplatzierten Fensters im Pool Spanien identisch. Abzüglich dieser Duplikate ergibt sich eine Gesamtzahl von 54 verschiedenen Genen in den Top 5 Kandidatenregionen der vier Pools. Unter den assoziierten *GO-Terms* dominieren unspezifische funktionelle Gruppen und Begriffe wie „*catalytic activity*“

Tabelle 28: Gene der fünf Kandidatenregionen mit den niedrigsten ZH_P-Werten der einzelnen Pools **(Fortsetzung auf der folgenden Seite)**

		ZH_P	Chr.	Gene	Produkt	Relevante *GO-Terms*
Ketsch	1	-3,09	11	GSVIVG01015456001	PDR-like ABC transporter	
				GSVIVG01015458001	WD-40 repeat protein	
				GSVIVG01015459001	Transporter, Sodium/bile acid symporter	
	2	-3,09	7	GSVIVG01003415001	Rab11/RabA-family small GTPase	
				GSVIVG01003416001	ribosomal RNA small subunit methyltransferase F	
				GSVIVG01003417001	putative peroxidase / Lignin-forming anionic peroxidase	
				GSVIVG01003418001	ribosomal protein S6	
				GSVIVG01003419001	serine/threonine protein kinase / Abscisic acid-inducible kinase	
	3	-2,94	18	GSVIVG01009365001	transcription initiation factor, putative / TFIID subunit 4B	
				GSVIVG01009366001	casein kinase II alpha subunit	reproduction, inflorescence development
				GSVIVG01009367001	unknown, uncharacterised protein family SERF	
				GSVIVG01009368001	unknown	
				GSVIVG01009369001	Acyl-CoA-binding domain-containing protein	defense response, … to bacterium, … to fungus
	4	-2,92	14	GSVIVG01036253001	pentatricopeptide repeat protein	
				GSVIVG01036252001	unknown	
	5	-2,90	7	GSVIVG01003420001	lanthionine synthetase C family protein, putative	
Pisa	1	-3,85	R	GSVIVG01006882001	unknown	pollen exine formation
				GSVIVG01006883001	unknown	
	2	-3,63	9	GSVIVG01034254001	unknown	
	3	-3,54	18	GSVIVG01034572001	histone-lysine N-methyltransferase	
	4	-3,51	18	GSVIVG01034571001	SET domain protein	
	5	-3,50	13	GSVIVG01016242001	Anion-transporting ATPase / ATPase GET3	
				GSVIVG01016240001	white-brown-complex ABC transporter family	
				GSVIVG01016239001	serine/threonine protein kinase	
Frankreich	1	-3,84	14	GSVIVG01023433001	60S ribosomal protein L18	
				GSVIVG01023432001	putative RPS2 / Disease resistance protein	defense response
	2	-3,60	8	GSVIVG01025745001	serine/threonine protein kinase / CDPK-related kinase	heat acclimation
				GSVIVG01025744001	unknown	
				GSVIVG01025743001	dual specificity phosphatase 12	
				GSVIVG01025742001	unknown	

		ZH_P	Chr.	Gene	Produkt	Relevante *GO-Terms*
Frankreich	2	-3,60	8	GSVIVG01025741001	phosphoglycerate dehydrogenase like 1	
	3	-3,52	19	GSVIVG01014072001	unknown / transducer and activator of transcription	
				GSVIVG01014073001	putative purine permease	
				GSVIVG01014074001	light harvesting chlorophyll a/b-binding protein	
				GSVIVG01014075001	methionine-R-sulfoxide reductase	
				GSVIVG01014076001	unknown / uncharacterized sugar kinase AF_0356	
	4	-3,50	R	GSVIVG01003477001	HAD-superfamily hydrolase, subfamily IIA	
				GSVIVG01003478001	serine acetyltransferase	
				GSVIVG01003479001	histone acetyltransferase complex component / Transcriptional adapter ADA2b	cold acclimation, response to cold
				GSVIVG01003480001	unknown	
				GSVIVG01003481001	agenet - & BAH-domain-containing protein	
				GSVIVG01003482001	unknown	
				GSVIVG01003483001	unknown	anther development, response to heat
	5	-3,47	8	GSVIVG01022506001	SPX- & zinc finger (C3HC4-type RING finger)-domain-containing protein	
				GSVIVG01022507001	FAD linked oxidase domain protein / D-lactate dehydrogenase	
Spanien	1	-3,76	10	GSVIVG01021482001	squalene-hopene cyclase	
				GSVIVG01021483001	beta-amyrin synthase	
				GSVIVG01021484001	beta-amyrin synthase	
	2	-3,55	19	GSVIVG01014072001	unknown / transducer and activator of transcription	
				GSVIVG01014073001	putative purine permease	
				GSVIVG01014074001	light harvesting chlorophyll a/b-binding protein	
				GSVIVG01014075001	methionine-R-sulfoxide reductase	
				GSVIVG01014076001	unknown / uncharacterized sugar kinase AF_0356	
	3	-3,47	10	vgl. Top 1		
	4	-3,43	5	GSVIVG01017617001	DUF814 domain protein, putative	
				GSVIVG01017620001	unknown	
				GSVIVG01017621001	unknown	
				GSVIVG01017622001	unknown	
				GSVIVG01017619001	aspartyl aminopeptidase	
	5	-3,40	10	GSVIVG01021485001	beta-amyrin synthase	

(28 Gene), „*binding*“ (27 Gene) oder „*metabolic process*“ (21 Gene). Daher wurden in Tabelle 28 nur Ontologien aufgenommen, die vor dem Hintergrund der Wilden Weinrebe als genetische Ressource für die Rebenzüchtung relevant erschienen.

Hierzu zählen primär *GO-Terms*, die pflanzliche Abwehrmechanismen gegenüber Pathogenen und Krankheiten beschreiben. In den Pools Ketsch und Frankreich ließ sich in diesem Zusammenhang jeweils ein Kandidatengen identifizieren. Es handelt sich um die Gene GSVIVG01009369001 (*Acyl-CoA-binding domain-containing protein*) und GSVIVG01023432001 (*putative RPS2 / disease resistance protein*). Ersteres befindet sich im drittplatzierten Fenster des Pools Ketsch (ZH_P=-2,94), in dem insgesamt fünf Gene annotiert sind. Das resultierende Protein ist 301 Aminosäuren (AS) lang und besitzt eine konservierte Bindedomäne für Acyl-Coenzym A. Als bestes Ortholog identifiziert das Online-Tool PLAZA das At4g24230-Gen (*ACBP3*) von *Arabidopsis thaliana*. Eine BLASTp-Suche führt zu verwandten Loci u.a. in der Walnuss (*Juglans regia*), der Orange (*Citrus sinensis*) oder dem Wunderbaum (*Ricinus communis*). Das zweite Gen ist im erstplatzierten Fenster des Pools Frankreich annotiert, das heißt der zugrundeliegende genetische Abschnitt wies in den vorangegangenen Analysen den genomweit niedrigsten ZH_P-Wert auf (ZH_P=-3,84). Neben dem angesprochenen GSVIVG01023432001-Locus ist nur ein weiteres Gen innerhalb des Fensters annotiert. Die 807 AS lange abgeleitete Peptidsequenz von GSVIVG01023432001 enthält eine NB-ARC-Domäne sowie Leucin-reiche Repeats (LRR). Sequenzverwandte Proteine lassen sich in verschiedenen Pflanzenspezies wie beispielsweise der Euphrat-Pappel (*Populus euphratica*), dem Maniok (*Manihot esculenta*) oder dem Kakaobaum (*Theobroma cacao*) nachweisen. Das Online-Tool PLAZA identifiziert kein konkretes Homolog in *Arabidopsis thaliana*, jedoch führen BLASTp-Suchen im *Arabidopsis*-Genom zu signifikanten Treffern bei zahlreichen NB-ARC- und LRR-enthaltenden Proteinen, denen ebenfalls eine Beteiligung in Abwehrmechanismen zugeschrieben wird.

Darüber hinaus sind Gene mit Bezug zu Fortpflanzungsprozessen von besonderem Interesse. Aus dieser Kategorie beinhaltet die Top 5-Liste die drei Kandidatengene GSVIVG01009366001 (*casein kinase II alpha subunit*), GSVIVG01034254001 (*unknown*) und GSVIVG01003483001 (*unknown*). Das erste der drei Gene befindet sich gemeinsam mit dem im Zuge der pflanzlichen Abwehrmechanismen besprochenen Gen für ein Protein mit Acyl-Coenzym A-Bindedomäne und drei weiteren Genen im drittplatzierten Fenster des Pools Ketsch (ZH_P=-2,94). Das resultierende 334 AS lange Protein enthält konservierte Domänen,

die es als eine Serin/Threonin-Kinase kennzeichnen. Assoziierte *GO-Terms* lassen auf eine Funktion in der Entwicklung des Blütenstands schließen. Homologe finden sich mit dem At3g50000-Locus in *Arabidopsis thaliana* sowie u.a. in der Zuckermelone (*Cucumis melo*), der Ölpalme (*Elaeis guineensis*) und im Sesam (*Sesamum indicum*). Das zweite mit Fortpflanzungsprozessen zusammenhängende Kandidatengen ist im zweitplatzierten Fenster des Pools Pisa annotiert (ZH_P=-3,63). Neben dem angesprochenen Gen GSVIVG01034254001 ist kein weiteres Gen in dem zugrundeliegenden genomischen Abschnitt beschrieben. Die abgeleitete 1.178 AS lange Proteinsequenz enthält trotz ihrer Länge keine beschriebenen konservierten Domänen. Zwar lässt sich mit dem Gen At5g58100 ein Homolog in *Arabidopsis thaliana* identifizieren, doch dieses kodiert genau wie die BLAST-Treffer aus der Purgiernuss (*Jatropha curcas*), dem Kakaobaum (*Theobroma cacao*) oder dem Pfirsich (*Prunus persica*) für ein uncharakterisiertes beziehungsweise hypothetisches Proteinprodukt. Diesem wird aber zumindest in *Arabidopsis thaliana* eine entscheidende Beteiligung bei der Bildung der Wand des Pollenkorns (Exine) zugeschrieben (Dobritsa et al. 2011). Bei dem letzten der drei Kandidatengene handelt es sich um GSVIVG01003483001, das aufgrund seiner Position im viertplatzieren Fenster des Pools Frankreich identifiziert wurde (ZH_P=-3,50). Dieser Locus enthält die größte Anzahl an Genen der Top 5-Liste. Neben dem angesprochenen Gen GSVIVG01003483001 sind noch sechs weitere Gene innerhalb des Fensters annotiert. Das resultierende 250 AS lange Proteinprodukt wird im Referenzgenom der Weinrebe als unbekannt angegeben und enthält darüber hinaus keine bekannten konservierten Domänen. Jedoch lassen sich bei *Arabidopsis thaliana* mit At1g32583 und At4G24972 zwei putative Homologe identifizieren, die über die Funktion des Gens Aufschluss geben können. Das letztere kodiert für TPD1 (tapetum determinant 1), einen Cystein-reichen Proteinliganden, für den eine Rolle in der Entwicklung der Antheren angenommen wird (Huang et al. 2016).

Weiterhin wurden Kandidatengene betrachtet, die Anpassungsreaktionen auf sich verändernde Umweltfaktoren vermitteln. In diesem Zusammenhang ließen sich in den fünf bestplatzierten Fenstern im Pool Frankreich insgesamt drei Gene identifizieren. Zwei dieser Gene sind mit der Akklimatisation an höhere Temperaturen oder Hitze assoziiert. Das erstere dieser beiden Gene, GSVIVG01025745001, ist im zweitplatzieren Fenster des Pools Frankreich annotiert (ZH_P=-3,60). Die abgeleitete Proteinsequenz ist 594 AS lang und enthält die charakteristischen Domänen einer Serin/Threonin-Kinase, die darüber hinaus in der Lage ist Ca^{2+}/Calmodulin zu binden. Das vom Online-Tool PLAZA identifizierte Homolog

aus *Arabidopsis thaliana*, At2g41140, ist an der Signaltransduktion nach einem Hitzeschock beteiligt und eine Überexpression des Gens vermittelt eine gesteigerte Thermotoleranz (Liu et al. 2008). Beim zweiten Gen handelt es sich um GSVIVG01003483001, das bereits im Zuge seiner möglichen Beteiligung an der Antherenentwicklung analysiert wurde. Es befindet sich mit sechs weiteren Genen im viertplatzierten Fenster des Pools Frankreich (ZH_P=-3,50). Wie schon erwähnt wurde neben dem besprochenen TPD1-Homolog mit dem Gen At1g32583 ein weiteres putatives Homolog in *Arabidopsis thaliana* identifiziert. Von dessen ersten Intron des 5'-UTRs wird eine microRNA (miRNA) transkribiert, die bei Hitzestress alternativ gespleißt wird (Yan et al. 2012). In der vorliegenden Version des Referenzgenoms der Weinrebe sind für das Gen GSVIVG01003483001 keine UTRs annotiert. Eine Suche nach einer mit der miRNA verwandten Sequenz im 5'-Bereich stromaufwärts der Genannotation ergab jedoch keine signifikanten Treffer. Ebenfalls im viertplatzierten Fenster des Pools Frankreich ist das dritte der ausgewählten Kandidatengene annotiert. Es handelt sich um das Gen GSVIVG01003479001 das für ADA2b, ein Bestandteil eines größeren Histonacetyltransferase-Komplexes, kodiert. Das resultierende Protein ist 540 AS lang und besitzt eine Vielzahl konservierter Domänen, darunter einen Zink-Finger des ZZ-Typs und eine Myb/SANT-ähnliche DNA-Bindedomäne. Eine BLASTp-Suche identifiziert verwandte Loci u.a. in der Walnuss (*Juglans regia*), der Chinesischen Dattel (*Ziziphus jujuba*) oder dem Wunderbaum (*Ricinus communis*). Das ADA2b-Homolog At4g16420 aktiviert in *Arabidopsis thaliana* die Expression von *cold-regulated* (COR)-Genen während der Akklimatisation an kältere Temperaturen (Pavangadkar et al. 2010).

Analyse der 33 Kerngene (*core genes*)

Als Kerngene (*core genes*) wurden, wie eingangs erwähnt, diejenigen Gene betrachtet, die in allen vier Pools als Kandidatengene auftraten und somit von besonderem Interesse waren. In Tabelle 29 sind die 33 Kerngene mit ihrer jeweiligen Gen-ID, dem ZH_P-Wert der zugrundeliegenden Fenster in den einzelnen Pools sowie dem resultierenden Proteinprodukt aufgelistet. Mehr als die Hälfte, nämlich 19 dieser 33 Gene, wurden bereits in der Top 5-Analyse identifiziert. Sie gehören demnach nicht nur zur Schnittmenge aller vier Pools, sondern sind gleichzeitig in den fünf Kandidatenregionen mit den niedrigsten ZH_P-Werten einzelner Pools lokalisiert. Bildet man in diesem Zusammenhang den Durchschnitt der ZH_P-Werte der 33 Kerngene in den einzelnen Pools und vergleicht diese

mit den durchschnittlichen ZH_P-Werten, die in der Gesamtheit der Kandidatenregionen dieser Pools gemessen wurden, so fällt auf, dass die Kerngene signifikant niedrigere ZH_P-Werte aufweisen. Am deutlichsten wird dieser Unterschied im Pool Spanien. Hier beträgt der durchschnittliche ZH_P-Wert der Kerngene -2,79, wohingegen der durchschnittliche ZH_P-Wert aller Kandidatenregionen dieses Pools nur -2,09 erreicht.

Keines der 33 Kerngene steht in Zusammhang mit pflanzlichen Abwehrmechanismen gegenüber Pathogenen und Krankheiten. Mit GSVIVG01003476001 und GSVIVG01036258001 konnten jedoch zwei gemeinsame Kandidatengene aller vier Pools identifiziert werden, die eine Rolle bei der Pollenentwicklung spielen und somit einen Bezug zu Fortpflanzungsprozessen aufweisen. Das erstere der beiden Gene kodiert für ein 1031 AS langes Protein, das neben zwei namensgebenden C2-Domänen eine GRAM-Domäne und zwei Kopien einer konservierten Domäne unbekannter Funktion (DUF4782) besitzt. Homologe finden sich mit dem At1g03370-Locus in *Arabidopsis thaliana* sowie u. a. im Kakaobaum (*Theobroma cacao*), in der Walnuss (*Juglans regia*) und im Maniok (*Manihot esculenta*). Beim zweiten Kandidaten handelt es sich um ein Gen für eine Phosphoribosylaminoimidazol-Carboxylase. Die abgeleitete 597 AS lange Proteinsequenz enthält dementsprechend die beiden typischen Domänen ATP-grasp und AIRC. Eine BLASTp-Suche identifiziert verwandte Loci u. a. im Kakaobaum (*Theobroma cacao*), im Zwergkrug (*Cephalotus follicularis*) und im Apfel (*Malus domestica*). Funktionsanalysen durch *loss-of-function*-Mutagenese des *Arabidopsis thaliana* Homologs At2g37690 zeigten Störungen in der Pollenentwicklung (Boavida et al. 2009). Darüber hinaus wurde eines der 33 Kerngene mit der Vermittlung von Thermotoleranz in Verbindung gebracht. Es handelt sich dabei um das bereits im Zuge der Top 5-Analyse besprochenene Gen GSVIVG01003479001, dessen Proteinprodukt ADA2b zumindest in *Arabidopsis thaliana* die Akklimatisation an kältere Temperaturen reguliert.

Tabelle 29: Gemeinsame Kandidatengene aller vier Pools

Gen	ZH_p				Produkt	Relevante *GO-Terms*
	Ke	Pi	Fr	Sp		
GSVIVG01003416001	-3,07	-3,10	-2,67	-2,54	ribosomal RNA small subunit methyltransferase F	
GSVIVG01003417001	-3,07	-3,10	-2,67	-2,54	putative peroxidase / Lignin-forming anionic peroxidase	
GSVIVG01003418001	-3,07	-3,10	-2,67	-2,54	ribosomal protein S6	
GSVIVG01003419001	-3,07	-3,10	-2,67	-2,50	serine/threonine protein kinase	
GSVIVG01003420001	-3,07	-3,03	-2,67	-2,50	lanthionine synthetase C family protein, putative	
GSVIVG01003421001	-2,30	-2,97	-2,67	-2,40	14-3-3 protein	
GSVIVG01003422001	-2,30	-2,97	-2,67	-2,21	pentatricopeptide (PPR) repeat-containing protein	
GSVIVG01003423001	-2,30	-2,97	-2,67	-2,12	histone H3	
GSVIVG01003424001	-2,30	-2,97	-2,67	-1,91	signal transducer, putative	
GSVIVG01003476001	-2,54	-2,44	-3,33	-2,39	putative C2 domain-containing protein	pollen maturation
GSVIVG01003477001	-2,54	-2,44	-3,50	-2,80	HAD-superfamily hydrolase, subfamily IIA	
GSVIVG01003478001	-2,69	-2,44	-3,50	-2,80	serine acetyltransferase	
GSVIVG01003479001	-2,77	-2,44	-3,50	-2,80	histone acetyltransferase complex component	cold acclimation, response to cold,
GSVIVG01003480001	-2,77	-2,44	-3,50	-2,80	unknown	
GSVIVG01003481001	-2,77	-2,42	-3,50	-2,80	agenet - & BAH-domain-containing protein	
GSVIVG01003482001	-2,77	-2,42	-3,50	-2,80	unknown	
GSVIVG01003950001	-2,43	-2,48	-2,38	-2,76	unknown	
GSVIVG01003951001	-2,63	-2,60	-2,38	-2,76	unknown	
GSVIVG01003952001	-2,63	-2,60	-2,38	-2,76	unknown	
GSVIVG01003953001	-2,63	-2,60	-2,38	-2,76	exostosin family protein	
GSVIVG01014071001	-2,43	-2,38	-3,09	-3,22	signal peptidase I	
GSVIVG01014072001	-2,43	-2,67	-3,52	-3,55	unknown / transducer and activator of transcription	
GSVIVG01014073001	-2,43	-2,67	-3,52	-3,55	putative purine permease	
GSVIVG01014074001	-2,43	-2,67	-3,52	-3,55	light harvesting chlorophyll a/b-binding protein	
GSVIVG01014075001	-2,43	-2,67	-3,52	-3,55	methionine-R-sulfoxide reductase	
GSVIVG01014076001	-2,43	-2,67	-3,52	-3,55	unknown /uncharacterized sugar kinase AF_0356	
GSVIVG01021482001	-2,32	-2,69	-2,60	-3,76	squalene-hopene cyclase	
GSVIVG01021483001	-2,32	-2,69	-2,60	-3,76	beta-amyrin synthase	
GSVIVG01021484001	-2,32	-2,69	-2,60	-3,76	beta-amyrin synthase	
GSVIVG01026025001	-2,53	-2,53	-2,28	-1,90	pentatricopeptide repeat-containing protein-like	
GSVIVG01036255001	-2,50	-2,48	-2,50	-2,47	pentatricopeptide repeat protein	
GSVIVG01036257001	-2,29	-2,48	-2,25	-2,05	unknown	
GSVIVG01036258001	-2,29	-2,61	-2,25	-2,05	Phosphoribosylamino-imidazole carboxylase	pollen development

Für die drei Kandidatengene GSVIVG01003476001, GSVIVG01003479001 und GSVIVG01036258001 wurde überprüft, ob in den Wilden Weinreben im Vergleich zum Referenzgenom der Kulturrebe Pinot Noir fixierte Varianten vorliegen, da es sich dabei um die ursprünglichen Angriffspunkte des Selektionsereignis handeln könnte. Als fixierte Varianten wurden Positionen betrachtet, an denen mehr als 90 % aller Wildreben-*reads* ein alternatives, nicht im Referenzgenom vorhandenes Allel zeigen. Insgesamt konnten sieben derartige Polymorphismen identifiziert werden, sechs davon befinden sich im Gen GSVIVG01003479001, einer im Gen GSVIVG01003476001 (Tabelle 30). Im Kandidatengen GSVIVG01036258001 liegen keine fixierten Varianten vor. Drei der sieben Polymorphismen befinden sich in proteinkodierenden Abschnitten, davon verursachen wiederum zwei einen Aminosäureaustausch im resultierenden Proteinprodukt. Der erste nicht-synonyme Nukleotidaustausch betrifft das zweite Exon des Gens GSVIVG01003476001. Infolgedessen liegt an Position 283 des abgeleiteten Proteins im Bereich der Domäne unbekannter Funktion (DUF4782) anstelle der im Pinot Noir vorhandenen aromatischen Aminosäure Phenylalanin (Phe) in den Wilden Weinreben das aliphatische Leucin (Leu) vor. Die zweite nicht-synonyme fixierte Variante liegt im 14. und zugleich letzten Exon des für die Phosphoribosylaminoimidazol-Carboxylase kodierenden Gens GSVIVG01003479001. Hier ersetzt an Position 508 des Polypeptids in den Wilden Weinreben ein Threonin (Thr) das Serin (Ser) der Kulturrebe. Zwar befindet sich damit der Aminosäureaustausch in der für die Katalyse der Carboxylierungs-Reaktion essentiellen AIRC-Domäne, jedoch handelt es sich sowohl bei Threonin als auch bei Serin um acyclische, polare Aminosäuren, so dass die Auswirkungen auf die Proteinstruktur vermutlich gering sind.

Tabelle 30: Fixierte Varianten in ausgewählten Kerngenen der Wilden Weinreben

Gen	Position	Frequenz des Referenzallels	Frequenz des alternativen Allels	AS-Austausch
GSVIVG01003476001	1410 (Exon)	A = 1,3 %	C = 98,7 %	Phe 283 Leu
GSVIVG01003479001	2263 (Exon)	A = 0 %	G = 100 %	
	2359 (Intron)	G = 0 %	A = 100 %	
	2650 (Intron)	G = 0 %	A = 100 %	
	2766^2767 (Intron)	- = 3,7 %	AG = 96,3 %	
	4221 (Intron)	C = 6,25 %	G = 93,75 %	
	4853 (Exon)	G = 1,7 %	C = 98,3 %	Ser 508 Thr

Analyse statistisch angereicherter *GO-Terms*

Den 2.601 Kandidatengenen konnten wie eingangs erwähnt insgesamt 3.461 *GO-Terms* zugeordnet werden. Diese verteilen sich auf die drei übergeordneten Bereiche „Biologischer Prozess“ (2.236 GO-Terms), „Molekulare Funktion“ (877 GO-Terms) sowie „Zelluläre Komponente“ (348 GO-Terms). Letztere wurden in den folgenden Analysen nicht weiter berücksichtigt. In der hierarchischen Struktur der Genontologie wurden den Kandidatengenen in den nächsten Instanzen weitere, speziellere Attribute zugeteilt. Ein Großteil der Kandidatengene steht hierbei im Zusammenhang mit allgemeinen Stoffwechselvorgängen innerhalb einer Zelle wie „*binding*“ (1.248 Gene), „*metabolic process*“ (996 Gene), „*cellular process*“ (949 Gene) oder „*catalytic activity*“ (871 Gene). Diese und weitere 124 *GO-Terms* sind mit 100 oder mehr Genen assoziiert. Fast die Hälfte der *GO-Terms* sind hingegen nur einem einzigen der Kandidatengene zugeordnet. Hier finden sich spezifischere Funktionen und Prozesse wie beispielsweise „*inositol phosphate dephosphorylation*“, „*inflorescence development*“ oder „*lateral root formation*“.

Die Aussagekraft dieser Informationen ist jedoch ohne *a priori* Wissen über die Frequenz der jeweiligen *GO-Terms* im Gesamtgenom nur gering. Daher wurde im Zuge eines sogenannten *GO-Enrichments* überprüft, welche *GO-Terms* in den Kandidatengenen im Vergleich zum Gesamtgenom statistisch angereichert ($p \leq 0{,}01$) und somit überrepräsentiert sind. Abbildung 27 zeigt die 54 *GO-Terms* aus den Bereichen „Biologischer Prozess“ und „Molekulare Funktion“, die in den 2.601 Kandidatengenen vor dem Hintergrund des Gesamtgenoms überrepräsentiert sind. Drei funktionelle Gruppen stechen hierbei besonders hervor. Zum einen sind dies Ontologien im Zusammenhang mit der Prozessierung des 3‘-Endes von tRNAs und anderen nicht-kodierenden RNAs („*tRNA 3‘-end processing*“: 10,2-fache Anreicherung, „*ncRNA 3‘-end processing*“: 7,3-fache Anreicherung, „*3‘-tRNA processing endoribonuclease activity*“: 10,2-fache Anreicherung). Eine vergleichbare Dominanz zeigen Gengruppen, die dem Transport von Kupferionen zugeordnet werden können („*copper ion transmembrane transport*“: 7,3-fache Anreicherung, „*copper ion transport*“: 5,1-fache Anreicherung, „*copper ion transmembrane transporter activity*“: 5,1-fache Anreicherung). Bei der dritten überrepräsentierten funktionellen Gruppe handelt es sich um Transportprozesse insbesondere von Proteinen in die Vakuole („*protein targeting to vacuole*“: 3,2-fache Anreicherung, „*establishment of protein localization to vacuole*“: 3,2-fache Anreicherung, „*protein localization to vacuole*“:

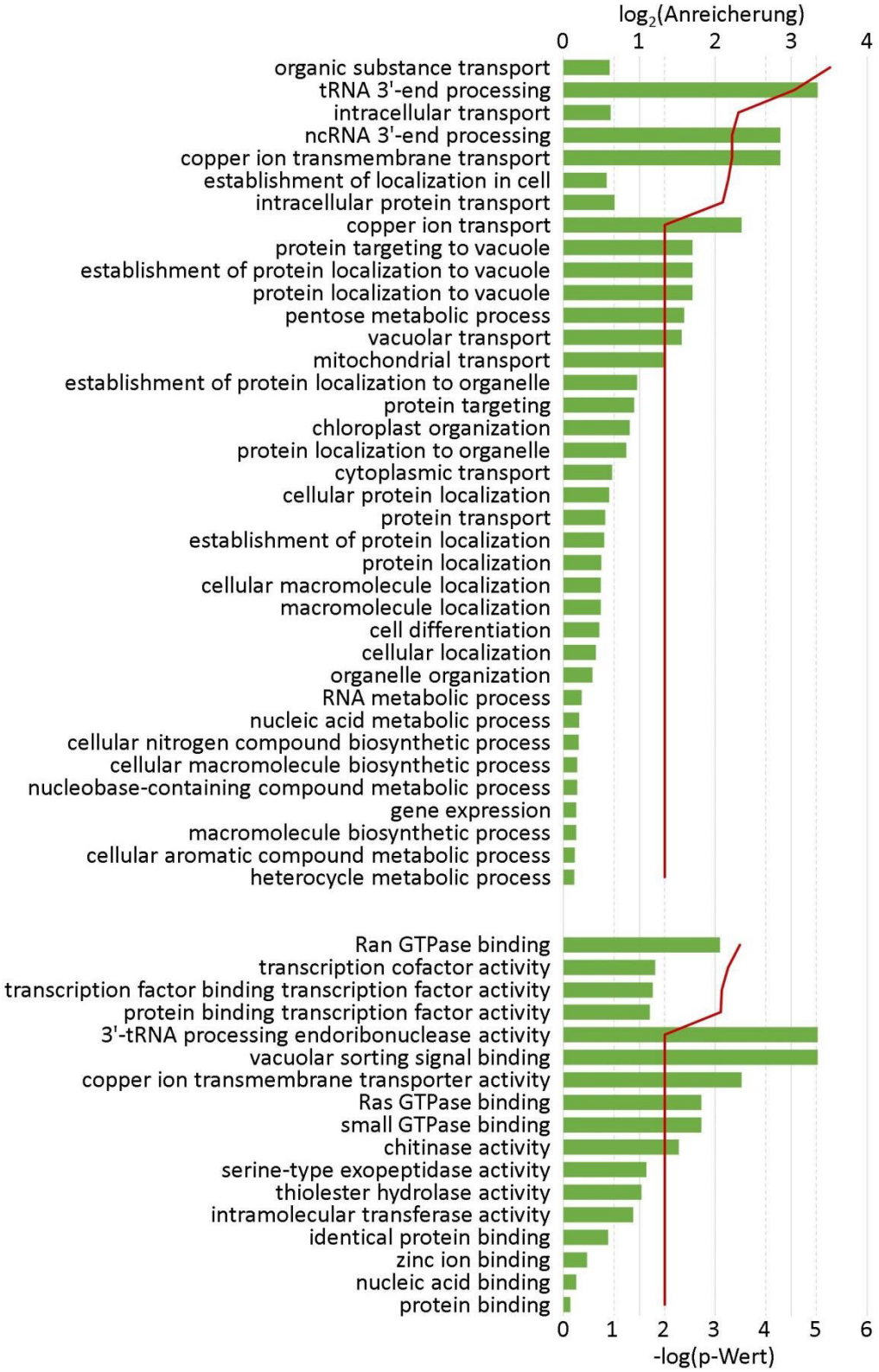

Abbildung 27: Statistisch angereicherte *GO-Terms* in den Kandidatengenen

Dargestellt sind *GO-Terms* aus den Bereichen „Biologischer Prozess“ (oben) und „Molekulare Funktion“ (unten), die in den 2.601 Kandidatengenen im Vergleich zum Gesamtgenom statistisch angereichert sind ($p \leq 0{,}01$). Die Anreicherung ist in Form einer $\log_2$-Skala angegeben, ein Wert von 3 entspricht demnach einer 8-fachen Anreicherung ($2^3=8$). Die rote Linie gibt die Signifikanz der Anreichung in Form des p-Werts an.

3,2-fache Anreicherung, „*vacuolar transport*“: 2,9-fache Anreicherung, „*vacuolar sorting signal binding*“ 10,2-fache Anreicherung).

Analyse Züchtungs-relevanter *GO-Terms*

Die Wilde Weinrebe *Vitis vinifera* subsp. *sylvestris* stellt als *crop wild relative* eine wichtige genetische Ressource für die Rebenzüchtung dar. Daher wurden Gengruppen, die für die Züchter von besonderem Interesse sind, identifiziert und die zugehörigen Kandidatengene gesondert analysiert (Tabelle 31, sowie Tabelle S12 im elektronischen Anhang). Hierzu gehörten 93 Gene mit Bezug zu pflanzlichen Abwehrmechanismen gegenüber Pathogenen und Krankheiten, die daher potenziell Resistenzen vermitteln können. Darüber hinaus wurden 125 Kandidatengene, die eine Rolle bei Fortpflanzungsprozessen spielen, identifiziert. Vor dem Hintergrund des Klimawandels wurden weiterhin Gene ausgewählt, die für Anpassungsreaktionen auf sich verändernde Umweltfaktoren wie Temperatur (47 Gene) oder Wasserverfügbarkeit (21 Gene) verantwortlich sind.

Tabelle 31: Züchtungs-relevante Kandidatengene

Züchtungs-relevanter *GO-Term*	Kandidaten-gene gesamt	Pool Ketsch	Pool Pisa	Pool Frank-reich	Pool Spanien
„Defense response“	93	19 (12)	17 (7)	29 (11)	51 (42)
„Reproduction“ & „gametophyte development“	125	48 (28)	41 (14)	50 (20)	41 (21)
„Response to temperature stimulus“	47	14 (8)	10 (3)	16 (7)	24 (15)
„Response to water deprivation“	21	7(4)	2 (1)	11 (5)	9 (4)

Werte in Klammern geben die Anzahl poolspezifischer Kandidatengene an.

Für diese Züchtungs-relevanten Kandidatengene wurden falls möglich mit dem Online-Tool PLAZA Homologe in *Arabidopsis thaliana* identifiziert, für die wiederum mithilfe der Datenbank TAIR eine Literaturrecherche durchgeführt wurde. Bei der Sichtung der Publikationen wurden gezielt Kandidatengene ausgewählt, deren Funktion im Modellorganismus *Arabidopsis thaliana* eine praktische Anwendung in der Rebenzüchtung sinnvoll erscheinen lässt. Als Beispiel sind an dieser Stelle Gene zu nennen, die in *Arabidopsis thaliana* eine Resistenz gegen Pathogene vermitteln, die auch die Kulturrebe befallen und somit ein Problem im Weinbau darstellen.

4 Diskussion

4.1 Authentifizierung Wilder Weinreben

Für Genomanalysen der Wilden Weinrebe *Vitis vinifera* subsp. *sylvestris* ist es notwendig, vorab die taxonomische Identität des zu sequenzierenden Materials eindeutig zu klären. Dies ist darin begründet, dass mit der Edlen Weinrebe *Vitis vinifera* subsp. *vinifera* eine Schwesternsubspezies existiert, die darüber hinaus in Europa deutlich weiter verbreitet ist als die vom Aussterben bedrohte Wilde Weinrebe. Zwar unterscheiden sich die beiden Subspezies in zahlreichen morphologischen und physiologischen Eigenschaften, wie beispielsweise der Beerenfarbe oder dem Geschlecht der Blüten, jedoch erschweren verschiedene Faktoren eine eindeutige Zuordnung einzelner Rebstöcke zu einer der beiden Subspezies. So wurden in den natürlichen Habitaten der Wilden Weinrebe verwilderte Kultur- und Unterlagsreben sowie durch Spontanbastardisierung entstandene Hybride zwischen Edlen und Wilden Weinreben identifiziert (Arroyo-García & Revilla 2013; Zecca et al. 2010; Bodor et al. 2010). Letztere entstanden vermutlich durch immigrierten Pollen aus benachbarten Weinbergen, dessen Einfluss je nach Habitat und analysiertem Jahr auf 4,2 bis 26 % geschätzt wird (Di Vecchi-Staraz et al. 2009). Liegt die Bastardisierung mit der Kulturform bereits mehrere Generationen zurück, wie es beispielsweise im Fall eines Individuums der Halbinsel Ketsch von Ledesma-Krist et al. (2015) beschrieben wurde, kommt es zu einer „Verwässerung“ des genetischen Einflusses des Kulturreben-Elternteils und Wildrebeneigenschaften treten wieder stärker hervor. Dies birgt jedoch bei Selektionsanalysen mittels Pool-Sequenzierung eine besondere Gefahr, da hier seltene Allelvarianten gezielt identifiziert und für die Analyse genutzt werden. Weiterhin wird die Zuordnung der Subspezies dadurch erschwert, dass einige der als Unterscheidungskriterium herangezogenen morphologischen Merkmale eine große Varianz oder Formenvielfalt zeigen. Daher eignen sich Eigenschaften wie beispielsweise die Samen- oder Blattform nur bedingt für die zweifelsfreie Unterscheidung der Subspezies (Rivera Núñez et al. 2007; Barth et al. 2009). Nicht zuletzt steht zudem ein Großteil der morphologischen Kennzeichen nur zeitlich sehr eingeschränkt zur Verfügung. Das Geschlecht der Weinreben lässt sich ausschließlich zur Blütezeit Anfang Juni bestimmen, die Beerenfarbe kann hingegen erst zum Zeitpunkt der Reife im Herbst ermittelt werden.

Der hohe Stellenwert einer sorgfältigen Authentifizierung des Wildreben-Materials konnte im Rahmen dieser Arbeit anhand der Probe L-17-12-2 deutlich gemacht werden. Diese wurde in Form von genomischer DNA von einem Kooperationspartner zur Verfügung gestellt, in der Annahme, dass es sich bei dem entsprechenden Individuum um eine weibliche Wilde Weinrebe aus der größten deutschen *sylvestris*-Population in den Ketscher Rheinauen handelt. Eine persönliche Inaugenscheinnahme des Geschlechts der Blüten auf der Anbaufläche des Kooperationspartners ergab jedoch, dass bei der Weinrebe L-17-12-2 sowohl die Staubblätter als auch das Pistill vollständig entwickelt sind, was vermuten lässt, dass es sich im vorliegenden Fall um hermaphroditische Blüten handelt. Laut Antcliff (1980) wird das Geschlecht in der Weinrebe durch einen einzelnen Locus mit drei Allelen festgelegt. In Kreuzungsversuchen konnte gezeigt werden, dass das männliche Allel hierbei dominant über das zwittrige Allel ist, welches wiederum Dominanz gegenüber dem weiblichen Allel zeigt (Dalbó et al. 2000; Antcliff 1980). Eine Marker-PCR zur Bestimmung des vorliegenden Genotyps der geschlechtsbestimmenden Region ergab, dass die Weinrebe L-17-12-2 heterozygot das weibliche F-Allel und das männlich-zwittrige MH-Allel trägt. Zwar erlauben die verwendeten Marker von Fechter et al. (2012) keine Differenzierung zwischen der männlichen und der hermaphroditischen Allelvariante, jedoch schließt das Vorhandensein des MH-Allels nach den bekannten Dominanzverhältnissen ein weibliches Geschlecht der analysierten Weinrebe aus. Die phänotypische Untersuchung der Blüten konnte somit auf genomischer Ebene verifiziert werden.

Als zweites morphologisches Kriterium zur Authentifizierung der Weinrebe L-17-12-2 wurde die Beerenfarbe zum Zeitpunkt der Reife betrachtet. Ähnlich wie bei der Untersuchung des Blütengeschlechts zeigte L-17-12-2 auch hier eine für Wilde Weinreben untypische Merkmalsausprägung in Form von rosa gefärbten Beeren, die eher eine Zuordnung zur Subspezies der Edlen Weinrebe nahelegen. Die der Beerenfarbe zugrundeliegende genomische Region umfasst zwei benachbarte Gene auf Chromosom 2, die für Transkriptionsfaktoren der MYB-Familie kodieren (Matus et al. 2008). Die beiden MYB-Transkriptionsfaktoren bilden mit weiteren Proteinen einen Komplex, der die Expression von Strukturgenen der Anthocyanbiosynthese positiv reguliert und somit die dunkle Farbe des Beerenexokarps verursacht (Koes et al. 2005). In Weinreben mit ungefärbten, d. h. „weißen“ Trauben, verhindert die homozygote Insertion eines *Gypsy*-Retrotransposons im Promotorbereich von *VvMybA1*, dem ersten der beiden angesprochenen Gene, die Expression dieses Gens (Kobayashi et al. 2004). *VvMybA2*, das zweite Gen, weist in

„weißen“ Weinreben eine homozygote Leserastermutation im dritten Exon auf, die ein unvollständiges, vermutlich defektes Proteinprodukt zur Folge hat (Walker et al. 2007). Mittels PCR und Sanger-Sequenzierung wurde im Rahmen dieser Arbeit nachgewiesen, dass beide homologe *VvMybA2*-Kopien die erwähnte Leserastermutation tragen. Darüber hinaus ist eine Kopie von *VvMybA1* durch die Insertion des Retrotransposons *Gret1* inaktiviert. Die zweite Kopie von *VvMybA1* scheint, soweit sich dies anhand von Datenbankabgleichen beurteilen ließ, intakt zu sein. Jedoch liegt im 5‘ Promotorbereich dieser Kopie ein 44 bp langes Indel vor. Möglicherweise beeinflusst diese zusätzlich vorhandene Sequenz innerhalb des Promotors die Expressionsstärke des *VvMybA1*-Gens in L-17-12-2 und verursacht so die beobachtete auffällig rosa Färbung der Beeren. Verschiedene Studien identifizierten eine ähnliche Insertion im Promotor des *VvMybA1*-Gens in anderen Weinreben und diskutieren ebenfalls eine mögliche Assoziation des Polymorphismus mit einer rötlichen oder rosa Beerenfarbe (This et al. 2007; Shimazaki et al. 2011).

Zusammengefasst rechtfertigen also beide im Rahmen dieser Arbeit untersuchten Merkmale sowohl auf genomischer Ebene als auch hinsichtlich des ausgeprägten Phänotyps eher eine Einordnung von L-17-12-2 als Edle Weinrebe anstelle der Wilden Weinrebe. Jedoch dokumentieren einige Feldstudien die Existenz vereinzelter Individuen, die ebensolche für *Vitis vinifera* subsp. *sylvestris* untypische Merkmale zeigten und trotzdem als echte Wilde Weinreben gelten (Ekhvaia & Akhalkatsi 2010; Anzani et al. 1990). Daher wurde zur endgültigen Aufklärung der Identität der Weinrebe L-17-12-2 ein genetischer Fingerabdruck mittels SSR-Markern angefertigt. SSR-Marker sind repetitive Sequenzen, die tandemartig im Genom angeordnet sind und aufgrund der 1 bis 6 bp kurzen Länge der Wiederholungseinheit als Mikrosatelliten bezeichnet werden (Ellegren 2004). Die Anzahl der Wiederholungen innerhalb eines Mikrosatelliten-Clusters ist aufgrund eines als „*replication slippage*“ bekannten Phänomens hochvariabel und lässt sich in Form eines Fragmentlängenpolymorphismus durch PCR und Gelelektrophorese sichtbar machen (Tautz 1989). Im Rahmen des europäischen Projekts *GrapeGen06* wurde ein einheitlicher Satz von insgesamt neun SSR-Markern zur Genotypisierung von Weinreben entwickelt und eine Datenbank mit den Profilen von über 3.000 Reben angelegt (Bacilieri & This 2007). Diese neun SSR-Marker gelten für die Unterscheidung von *Vitis vinifera* subsp. *vinifera* und *Vitis vinifera* subsp. *sylvestris* als informativ (Ledesma-Krist et al. 2015) und wurden daher für die Authentifizierung der Weinrebe L-17-12-2 verwendet. Dabei wurden für drei der neun Marker insgesamt vier verschiedene Allelgrößen identifiziert, die in dieser Form nicht bei

Vitis vinifera subsp. *sylvestris* existieren. Daher wurde ausgeschlossen, dass es sich bei L-17-12-2 um eine echte Wilde Weinrebe der Subspezies *Vitis vinifera* subsp. *sylvestris* handelt. Die verbleibenden sechs Marker wiesen hingegen sieben verschiedene Allelgrößen auf, die allesamt für *Vitis vinifera* subsp. *sylvestris* bekannt sind. In der Wildrebenpopulation der Insel Ketsch sind diese Allele ausnahmslos vorhanden, jedoch mit unterschiedlich starker Verbreitung[2]. Für die Marker VVS2 und VVMD27 wurden in der Weinrebe L-17-12-2 Allele im homozygoten Zustand identifiziert, die in gleicher Form bei 84 bzw. 92 % der Ketscher Wildreben zu finden sind. Die für die Marker VrZAG79, VVMD32 und VVMD25 in L-17-12-2 gemessenen Allelgrößen sind hingegen mit 2 bis 8 % deutlich weniger in der Wildrebenpopulation der Insel Ketsch verbreitet. Beim Marker VVMD5 lagen heterozygot zwei Allele vor, von denen das eine mit 30 % eine moderate, das andere mit 3 % eine geringe Frequenz in der Ketscher Population zeigt. Die Kombination aus sowohl *sylvestris*-typischen als auch *sylvestris*-fremden Allelen an den analysierten SSR-Loci führt gemeinsam mit den Ergebnissen der Untersuchung des Blütengeschlechts und der Beerenfarbe zu der Schlussfolgerung, dass es sich bei der Weinrebe L-17-12-2 um einen Hybrid aus Edler und Wilder Weinrebe handelt. Das Vorhandensein von Allelgrößen, die in der Ketscher Wildrebenpopulation weit verbreitet sind, spricht dafür, dass L-17-12-2 das Ergebnis einer Spontanbastardisierung durch in die Ketscher Rheinauen immigrierten Pollen von Kulturreben sein könnte. Vergleichbare Fälle wurden bereits von Di Vecchi-Staraz et al. (2009) in zwei Habitaten in Frankreich und von Ledesma-Krist et al. (2015) für die Insel Ketsch beschrieben.

Neben den SSR-Markern eigneten sich auch die im Rahmen dieser Arbeit getesteten Transposon-basierten Marker für eine Authentifizierung Wilder Weinreben. Mit drei Primern der IRAP- und iPBS-Methode konnten ausreichend polymorphe Banden generiert werden, um zweifelsfrei zwischen Edlen und Wilden Weinreben zu unterscheiden. Sie wurden genutzt, um die Zuordnung von jeweils vier Individuen der Pools Ketsch und Pisa zur Subspezies der Wilden Weinreben zu verifizieren. Für die verbleibenden Individuen der Pools wurde auf eine detaillierte Überprüfung der taxonomischen Identität, wie sie hier für die Probe L-17-12-2 beschrieben wurde, verzichtet. Jedoch wurde Material von Kooperationspartnern bezogen, bei dem eine vorangegangene Genotypisierung die

[2] Die hier zum Vergleich herangezogenen Allelfrequenzen der Ketscher Wildrebenpopulation wurden freundlicherweise von Dr. Erika Maul (Institut für Rebenzüchtung, Julius Kühn-Institut Geilweilerhof) zur Verfügung gestellt.

Zuordnung zur Subspezies der Wilden Weinrebe bestätigte. Lediglich für zwei Weinreben lag eine solche Information nicht vor. In diesen Fällen wurde erneut eine Untersuchung des Geschlechts anhand der Blütenmorphologie vorgenommen.

4.2 Genomanalysen mittels *Next-Generation Sequencing*

Next-Generation Sequencing-Technologien haben das Feld der Genomsequenzierungen in den vergangenen zehn Jahren revolutioniert. Zu Zeiten der konventionellen Sanger-Sequenzierung dauerte die Entschlüsselung des Genoms der Acker-Schmalwand *Arabidopsis thaliana* noch mehr als vier Jahre und kostete schätzungsweise 50 bis 70 Mio. US$ (TAGI 2000; Bevan et al. 1997). Dank neuer Hochdurchsatzverfahren mit stark reduziertem Zeit- und Kostenaufwand sind heute hingegen Sequenzierprojekte von Pflanzen im „Alleingang“, also abseits großer, internationaler Konsortien, möglich (Bolger et al. 2014). So konnten im Rahmen dieser Arbeit insgesamt acht Genomsequenzierungen mit der Illumina-Technologie des HiSeq 2000 und des Nachfolgers HiSeq 2500 durchgeführt werden. In vier Fällen handelte es sich dabei um Sequenzierungen von einzelnen Individuen, darunter zwei Wilde Weinreben, eine Edle Weinrebe und eine Hybridrebe, zum Zwecke von komparativen Genomanalysen. Mit den verbleibenden vier Sequenzierungen sollten hingegen populationsgenetische und evolutionsbiologische Aspekte beleuchtet werden, weswegen hier sogenannte Pool-Sequenzierungen (*Pool-Seq*) durchgeführt wurden, bei denen die genomische DNA von 10 bis 15 Individuen vorab vereint wurde. Insgesamt wurden im Zuge der acht Sequenzierungen mehr als 140 Gigabasen (Gb) Sequenzinformation erzeugt. Dies entspricht einer 295-fachen Genomabdeckung bei einer geschätzten Größe des Weinrebengenoms von 475 Mb (Lodhi & Reisch 1995).

Jedoch stand nicht die gesamte Sequenzinformation für die nachfolgenden Analysen zur Verfügung. Aufgrund höherer Fehlerraten der Illumina-Technologie im Vergleich zur konventionellen Sanger-Sequenzierung mussten die Sequenzdaten zunächst einem Filterprozess sowie einer Qualitätskontrolle unterzogen werden (Dohm et al. 2008). Besonders zu berücksichtigen waren dabei eine erhöhte Fehleranfälligkeit an den 3‘- und 5‘-Enden, eine Kontamination durch Adapter und undefinierte Nukleotide. Dieser Schritt sorgte für die notwendige Verbesserung der Qualität der Sequenzen, verringerte jedoch die Gesamtinformation der Daten um knapp 18 %, so dass insgesamt 115 Gb verblieben. Weiterhin konnte gezeigt werden, dass bis zu 20 % der Sequenzen nicht vom Kerngenom,

sondern von den Genomen der Chloroplasten (cpDNA) und Mitochondrien (mtDNA) stammen. Dies ist darauf zurückzuführen, dass bei der Isolierung der genomischen DNA die Organellen nicht entfernt wurden. In grünen Blättern, die das Ausgangsmaterial für die DNA-Extraktion darstellten, bildet cpDNA 12 bis 23 % der Gesamt-DNA einer Zelle (Boffey & Leech 1982; Lamppa et al. 1980; Scott & Possingham 1980). Eine pflanzliche Zelle beinhaltet darüber hinaus durchschnittlich 200 bis 600 Mitochondrien (Logan 2007). Hinzu kommt, dass Blütenpflanzen die größten bekannten mitochondrialen Genome besitzen (Kubo & Newton 2008). Im Fall der Weinrebe umfasst die mtDNA mehr als 770 kb (Goremykin et al. 2008).

Um die tatsächliche Abdeckung des Kerngenoms der Weinrebe durch die Illumina-Daten abzuschätzen, wurde eine Kartierung der Sequenzen gegen das publizierte Referenzgenom durchgeführt. Je nach Datensatz konnte eine 18- bis 33-fache Abdeckung erreicht werden. Aufgrund der zu kleinen Größe des Ausgangsdatensatzes zeigte nur die Wilde Weinrebe aus dem Kaukasus eine deutlich geringere Abdeckung von unter 3x. Der Grund für das unverhältnismäßig geringe Datenvolumen der Kaukasus-Probe war vermutlich eine zu geringe Clusterdichte bei der Sequenzierung. Der kritische Schritt ist hierbei die akurate Quantifizierung der DNA-Bibliothek. Wird die DNA-Konzentration überschätzt, resultiert dies in einer zu starken Verdünnung der Bibliothek und letztlich in der angesprochenen niedrigen Clusterdichte (Quail et al. 2008). Die Quantifizierung der DNA-Bibliotheken erfolgte mit dem Agilent 2100 Bioanalyzer in Kombination mit dem Qubit Fluorometer. Beide Methoden sind nur in der Lage, die DNA-Konzentration der Probe zu ermitteln. Sie überprüfen jedoch nicht, ob die DNA-Moleküle in der Bibliothek an beiden Enden Adapter tragen und somit Cluster generieren können. Hierfür wäre eine quantitative PCR (qPCR) notwendig gewesen (Buehler et al. 2010). Neuste *droplet digital* PCR (ddPCR)-Protokolle verzichten sogar auf die bei der qPCR nötigen Standards und erlauben eine absolute Quantifizierung bei gleichzeitiger Bestimmung der Integratgrößen der Bibliothek (Laurie et al. 2013).

Es stellt sich die Frage, welche Abdeckung bei einer Genomsequenzierung tatsächlich notwendig ist. Bisher wurden nur sehr wenige Genome ausschließlich auf Basis von NGS-Daten der zweiten Generation, zu der die Illumina-Technologie gehört, *de novo* assembliert. Das bekannteste Beispiel dürfte das Genom des Großen Pandas (*Ailuropoda melanoleuca*) sein, das 2010 von einem internationalen Konsortium publiziert wurde. Dafür wurden

Sequenzdaten generiert, die das Pandagenom 56-fach abdecken (Li et al. 2010). Diese Abdeckung wird von keinem der hier vorgestellten Datensätze erreicht und könnte allenfalls durch eine Kombination der Proben gewährleistet werden. Das dadurch entstehende Metagenom aus verschiedenen Subspezies würde jedoch die ohnehin starke Heterozygosität des Rebengenoms zusätzlich potenzieren. Trotz der geringen Abdeckung wurde eine *de novo* Assemblierung der einzelnen sequenzierten Rebengenome getestet. Dabei wurden je nach Datensatz zwischen 130.000 und 250.000 *contigs* erzeugt. Sie wiesen eine Gesamtlänge von 300 bis 350 Mb auf, womit die *de novo* Assemblierungen ca. 63 bis 73 % des Weinrebengenoms abdecken. Der GC-Gehalt sank durch die *de novo* Assemblierung von 33,5 bis 36,8 % der gefilterten Daten auf einheitliche 33 % in den *contigs*. Damit ist der GC-Gehalt der Assemblierung mit dem des Referenzgenoms identisch (Jaillon et al. 2007). Die verschiedenen Assemblierungen unterschieden sich hinsichtlich ihrer Qualität, die sich anhand der durchschnittlichen *contig*-Länge und des N50-Werts abschätzen lässt (Miller et al. 2010). Die Assemblierung der Wilden Weinrebe Hördt29 und der Hybridrebe zeigten hierbei die besten Werte mit durchschnittlichen *contig*-Längen über 2,5 kb und N50-Werten über 8 kb. Die Genome der Pool-Sequenzierungen sowie der Kulturrebe Weißer Heunisch ließen sich deutlich schlechter zusammensetzen. Im Fall des Weißen Heunischs ist als Ursache der schlechteren Assemblierung sicherlich die stärkere Heterozygosität von Kulturreben im Vergleich zu Wilden Weinreben zu nennen (Grassi et al. 2008). Aus demselben Grund erniedrigten Jaillon et al. (2007) künstlich die Heterozygosität vor der Sequenzierung der Kulturrebe Pinot Noir durch wiederholte Selbstung. Die Heterozygosität von Hördt29 ist als Vertreter der Wilden Weinrebe erniedrigt, was die Assemblierung der zugehörigen Sequenzdaten entsprechend vereinfacht. Für eine Hybridrebe würde man aufgrund der Abstammung von zwei verschiedenen Subspezies eine höhere Heterozygosität und somit schlechtere Assemblierung erwarten. Jedoch lagen bei der Genotypisierung der Hybridrebe L-17-12-2 sieben der neun SSR-Loci homozygot vor. Bei den Pool-Sequenzierungen stammten die zu assemblierenden Daten von 10 bis 15 Individuen, das Zusammensetzen eines solchen Metagenoms gestaltet sich im Vergleich zu den Genomen der Einzelindividuen aufgrund interindividueller Polymorphismen ungleich schwerer.

Trotz der geringeren Genomabdeckung überstiegen die N50-Werte der hier vorgestellten *de novo* Assemblierungen um ein Vielfaches die der ersten Assemblierungsrunde des Pandagenoms, bei der nur ein N50-Wert von 1.483 bp erreicht wurde (Li et al. 2010).

Jedoch konnten Li und Kollegen die assemblierten *contigs* zu sogenannten Gerüststrukturen (*scaffolds*) verknüpfen, da ihnen Positionsinformationen aus Bibliotheken mit größeren Integraten zur Verfügung standen. Dadurch generierten sie letzten Endes ein Pandagenom mit einem N50-Wert von 40 kb. Die Assemblierung der Weinrebengenome war hingegen an dieser Stelle beendet. Zur Erstellung von Gerüststrukturen wären weitere Positionsinformationen, wie sie die Sequenzierung von *mate-pair*-Bibliotheken liefern, oder längere *reads*, beispielsweise durch eine Sequenzierung mit den Geräten der Firma Pacific Biosciences, vonnöten. Eine vollständige *de novo* Assemblierung war aber nicht erforderlich. Die im Rahmen dieser Arbeit durchgeführten Analysen transposabler Elemente oder die Suche nach Fußspuren der Selektion lassen sich mit alternativen Strategien wie etwa Kartierungen oder BLAST-Suchen beantworten.

4.2.1 Transposable Elemente in Rebengenomen

Der prozentuale Anteil transposabler Elemente am Genom der Pflanzen ist extrem variabel, selbst zwischen einigen nah verwandten Spezies lassen sich signifikante Unterschiede feststellen (Bennetzen 2000). Das heute bekannte Spektrum reicht dabei von 3 % im Zwerg-Wasserschlauch (*Utricularia gibba*) bis zu 85 % im Mais (Ibarra-Laclette et al. 2013; Schnable et al. 2009). Aufgrund ihrer Fähigkeit zur Transposition gehören transposable Elemente zum variabelsten Teil eines Genoms (Lisch 2013). Daher ist klar, dass ihre Rolle in der Architektur und Evolution von Genomen weit über die der „eigennützigen DNA“ („selfish DNA“), die sich nur im Genom ihres Wirts vermehrt und ihm dabei keinen Nutzen bietet, hinausreicht (Hurst & Werren 2001). In verschiedenen Fällen nahmen transposable Elemente Einfluss auf die Domestizierung und Evolution von Kulturpflanzen. Als Beispiele dienen die phänotypischen Veränderungen des Mais bei der Kultivierung aus der Urform Teosinte oder die Entstehung des Kolumnarwachstums beim Apfel (Studer et al. 2011; Otto et al. 2014). Mit den im Rahmen dieser Arbeit durchgeführten Hochdurchsatz-sequenzierungen standen genomische Sequenzdaten verschiedener Wilder Weinreben sowie einer Kulturrebe zur Verfügung, um die Rolle transposabler Elemente in der Evolution und Domestizierung der Weinrebe zu analysieren. Ergänzt wurden die vorliegenden Daten durch weitere genomische Illumina-Sequenzen aus dem *Sequence Read Archive* (SRA) des NCBI. Dabei fiel die Wahl auf eine weitere zur Weinproduktion eingesetzte Edle Weinrebe, eine Tafeltraube und eine amerikanische *Vitis*-Spezies, die als Unterlagsrebe verwendet

wird. Durch diese Vielfalt konnten die Kultivierungsformen der Weinreben intensiver beleuchtet werden.

Prinzipiell ist eine Analyse transposabler Elemente auf Basis einer *de novo* Assemblierung eben dieser mittels NGS-Daten möglich (Zytnicki et al. 2014). Jedoch gestaltet sie sich aufgrund des repetitiven Charakters der Elemente als sehr schwierig und das Ergebnis ist letztlich nur eine Konsensussequenz, die sich aus verschiedenen Kopien einer Transposonfamilie zusammensetzt. Informationen zur Dynamik transposabler Elemente in den verschiedenen Genomen erhält man auf diese Art und Weise nicht. Daher wurde zur Analyse der transposablen Elemente in der Weinrebe eine andere Strategie gewählt. In Annahme, dass die Häufigkeit eines DNA-Elements direkt mit der Anzahl zugehöriger Sequenzen in einem Datensatz korreliert, wurden sogenannte „*read-count*"-Analysen durchgeführt (Magi et al. 2012). Die Kopienzahl der transposablen Elemente in den zehn untersuchten Rebengenomen wurde mit Hilfe der RPKM-Formel berechnet, die sonst in RNA-Seq-Studien zur Bestimmung der Genexpression eingesetzt wird (Mortazavi et al. 2008).

Insgesamt wurden zwischen 9.500 und 26.000 Kopien transposabler Elemente in den Genomen der zehn analysierten Weinreben identifiziert, was einer Größe von 32 bis 84 Mb entsprach. Die transposablen Elemente machten somit 6 bis 18 % des jeweiligen Gesamtgenoms aus. Diese Werte zeigen eine gute Übereinstimmung mit den Berechnungen von Jaillon et al. (2007), die in der Kulturrebe Pinot Noir je nach verwendeter Strategie[3] zur Identifizierung transposabler Elemente eine Gesamtlänge von 53 bzw. 83 Mb und einen prozentualen Anteil von 11 bzw. 17 % ergaben. Für das zweite publizierte Weinrebengenom werden mit 108,5 Mb transposabler Elemente jedoch höhere Werte angegeben (Velasco et al. 2007). Dies lässt darauf schließen, dass eine große Variabilität hinsichtlich der Kopienzahl transposabler Elemente in den Genomen der Weinreben existiert. Die Ergebnisse von Jaillon et al. (2007) deuten jedoch auch an, dass die tatsächlich gemessene Kopienzahl stark von der verwendeten Detektionsmethode abhängt, was einen Vergleich zwischen verschiedenen Publikationen erschwert. Klasse I Transposons, darunter vorwiegend die LTR-Retrotransposons, dominieren in allen zehn untersuchten

[3] Jaillon et al. (2007) verwendeten zwei verschiedene Strategien zur Identifizierung transposabler Elemente in dem von ihnen sequenzierten Genom der Kulturrebe Pinot Noir. Beide werden ausführlich im SI-Anhang ihrer Publikation beschrieben und sollen daher an dieser Stelle nicht genauer besprochen werden.

Rebengenomen deutlich über den Klasse II Transposons. Dies steht im Einklang mit Literaturangaben zur Prävalenz der Retrotransposons in Eukaryoten und insbesondere in Pflanzen (Bennetzen 2000). Auch die beobachteten Häufigkeitsverhältnisse der einzelnen Familien, wie beispielsweise die höhere Kopienzahl der *Gypsy*- gegenüber der *Copia*-Elemente oder das seltenere Auftreten von SINEs, decken sich mit der Literatur (Di Genova et al. 2014; Velasco et al. 2007).

Naito et al. (2006) konnten nachweisen, dass die Kultivierung von Reis mit einer drastischen Expansion transposabler Elemente einherging. Diese Tendenz war in den vorliegenden Daten der Weinreben kaum zu erkennen. Einzig die isolierte Betrachtung der Sequenzierung von Einzelindividuen zeigte eine 1,1- bis 1,7-fache Steigerung der Kopienzahl von *Gypsy*- und LINE-Transposons in den beiden Kulturreben Weißer Heunisch und Tannat im Vergleich zur Wildrebe Hördt29. Bei der Edlen Weinrebe Sultanina trat eine solche Expansion nicht auf. Die Domestizierung dieser Tafeltraube erfolgte jedoch unabhängig von den Kulturreben für die Weinproduktion (Di Genova et al. 2014; Aradhya et al. 2003). Bezieht man die Pool-Sequenzierungen in die Analyse mit ein, zeigt sich ein gegensätzliches Bild in Form einer Expansion transposabler Elemente in den untersuchten Populationen der Wildrebe. Dies könnte mit Hypothesen zur stress-induzierten Aktivierung transposabler Elemente in Verbindung stehen. Bereits Barbara McClintock, die Entdeckerin der transposablen Elemente, bezeichnete die von ihr im Mais untersuchten Transposons als „Kontrollelemente“. Sie vermutete, dass diese in einer vom Aussterben bedrohten oder andersartig gestressten Population als Quelle der Hypermutagenizität dienen können, durch die wiederum überlebensfähige Individuen entstehen (McClintock 1984). Die massive Zerstörung und Fragmentierung der natürlichen Habitate sowie Pathogene könnten genau solche Stessoren für die Wilde Weinrebe darstellen. Eine Aktivierung transposabler Elemente könnte daher die Reaktion der Rebenpopulation auf die sich verändernden Umweltbedingungen sein. Durch die vielfältigen Effekte, wie beispielsweise die Inaktivierung oder veränderte Regulation von Genen, epigenetische Veränderungen oder chromosomale Rearrangements, die transposable Elemente in einem Genom verursachen, können durch Zufall Wilde Weinreben entstehen, die besser an die gegebenen Bedingungen angepasst sind und so das Überleben der Population sichern (Casacuberta & González 2013). Die Kulturreben sind dank Pflege, Düngung und Behandlung mit Pflanzenschutzmitteln durch die Winzer deutlich weniger umweltbedingten Stressfaktoren ausgesetzt, so dass keine Notwendigkeit für die Kreation neuer Diversität durch die

Aktivierung transposabler Elemente besteht. Darüber hinaus erfolgt durch den Winzer eine strenge Selektion. Andersgeartete Phänotypen – wie sie möglicherweise durch die o. g. genomischen Veränderung infolge einer Aktivierung von transposablen Elementen entstehen könnten, werden konsequent von der Anbaufläche entfernt und nicht weiter vermehrt.

Allerdings ist zu beachten, dass dieser Expansionseffekt der transposablen Elemente auf die Pool-Sequenzierungen beschränkt ist. In der Einzelsequenzierung der Wilden Weinrebe Hördt29 lässt er sich nicht beobachten. Eine Erklärungsmöglichkeit für diese Unterschiede zwischen Pool- und Einzelsequenzierungen ist, dass die Varianz in der Kopienzahl selbst in der Subspezies der Wilden Weinrebe sehr hoch ist und mit der Probe Hördt29 durch Zufall ein Individuum mit niedriger, in den Populationen hingegen Individuen mit durchschnittlich höherer Kopienzahl ausgewählt wurden. Eine solch hohe intraspezifische Variabilität wurde bereits in der Sonnenblume beschrieben (Mascagni et al. 2015). Andererseits könnten die beobachteten Unterschiede auch technisch bedingt sein. Die Pool-Sequenzierungen wurden etwa ein bis zwei Jahre nach der Sequenzierung der Einzelindividuen durchgeführt. Veränderungen bei der Bibliothekserstellung und in der Sequenzierchemie könnten für eine verzerrte Repräsentation repetitiver Anteile des Genoms durch die Sequenzdaten der verschiedenen Proben verantwortlich sein (Poptsova et al. 2014). Zur Überprüfung der beobachteten Unterschiede existieren vielfältige molekularbiologische Methoden. Eine Fluoreszenz-*in-situ*-Hybridisierung, wie sie bei *Drosophila melanogaster* zur Charakterisierung verschiedener transposabler Elemente verwendet wurde, gestaltet sich bei den Weinreben aufgrund der kleinen Chromosomen schwierig, wenn auch nicht unmöglich (Zakharenko et al. 2007; Jiang & Gill 2006). Besser geeignet ist die Southern-Hybridisierung, die bereits erfolgreich zur Analyse von Kopienzahlunterschieden von SINE-Elementen bei Nachtschattengewächsen eingesetzt wurde (Wenke et al. 2011).

Die Weinrebe *Vitis rotundifolia* unterscheidet sich hinsichtlich der Kopienzahlen der transposablen Elemente signifikant von den anderen Proben, weswegen sie in den Analysen eine Art Außengruppe darstellt. Mit insgesamt 10.000 Kopien konnten nur etwa halb bis ein Drittel so viele transposable Elemente identifiziert werden wie in den verbleibenden Weinreben. Tatsächlich könnten im Genom der amerikanischen *Vitis*-Spezies deutlich weniger transposable Elemente vorhanden sein. Allerdings liegt die Vermutung nahe, dass die Ursache dieser beobachteten Unterschiede technischer Natur ist. Die angewendete

Strategie zur Identifizierung transposabler Elemente nutzte eine Kartierung gegen bekannte Transposonsequenzen der Weinrebe. Die verwendete Datenbank wurde primär aus Transposons von *Vitis vinifera* zusammengestellt. Zwar wurde die Stringenz der Kartierung mit 80 % bewusst niedrig gewählt, jedoch lässt sich nicht ausschließen, dass sie dennoch zu hoch für die vorliegende interspezifische Kartierung der Sequenzdaten war. Die amerikanischen *Vitis*-Spezies, zu denen *Vitis rotundifolia* zuzuordnen ist, trennten sich von den europäischen und asiatischen Vertretern der Gattung bereits vor ca. 6 bis 6,6 Mio. Jahren (Zecca et al. 2012). Transposable Elemente, die bereits im letzten gemeinsamen Vorfahren der amerikanischen und europäischen Weinreben vorhanden waren, zeigen heute in den beiden analysierten Spezies vermutlich eine hohe Sequenz-Divergenz. Hierdurch entsteht eine Kartierungsinkompatibilität der *rotundifolia*-Sequenzdaten gegen die *vinifera*-Referenz, die letztendlich für die Unterschätzung der Kopienzahl in *Vitis rotundifolia* verantwortlich ist. Eine solche Anfälligkeit der verwendeten Kartierungsstrategie bei interspezifischen Vergleichen wurde bereits in der Sonnenblume und im Kakaobaum beschrieben (Tetreault & Ungerer 2016; Sveinsson et al. 2013).

Aufgrund der beobachteten Kopienzahlunterschiede bietet sich eine Nutzung der transposablen Elemente als molekulare Marker für die Genotypisierung oder zur Erstellung von Kopplungskarten an (Seiwert 2014). Im Rahmen dieser Arbeit werden zwei Strategien vorgestellt, die die vorliegenden Insertionspolymorphismen transposabler Elemente in den verschiedenen Weinreben nutzen und auf ihrer Basis ein spezifisches genetisches Profil erzeugen. Die verwendeten Primer generierten eine ausreichende Anzahl polymorpher Amplifikate, um zwischen den beiden Subspezies *Vitis vinifera* subsp. *vinifera* und *Vitis vinifera* subsp. *sylvestris* zu unterscheiden, was sogleich für die Authentifizierung von acht Wilden Weinreben der Pool-Sequenzierung genutzt wurde (vgl. Kapitel 4.1). Für eine Differenzierung zwischen den im Rahmen dieser Arbeit analysierten Populationen Wilder Weinreben genügte die Auflösung jedoch nicht. Zwar gruppierten jeweils zwei bzw. drei der analysierten Individuen gemäß ihrer Herkunft, jedoch bildeten die verbleibenden Individuen aus Ketsch und Pisa ein gemeinsames Cluster. Für deren Trennung hätte es vermutlich weiterer Primer bedurft, um die Analyse auf zusätzliche transposable Elemente und somit weitere polymorphe Loci auszudehnen. Prinzipiell ist die Trennung der verschiedenen Wildreben-Populationen möglich, sie konnte bisher jedoch nur mit SSR-Markern gezeigt werden (De Andrés et al. 2012; Grassi et al. 2008). Für die Wildrebe aus dem Kaukasus konnten mit den verwendeten Transposon-basierten Markern nur sehr

wenige oder schwach ausgeprägte Amplifikate erzeugt werden. In der auf Basis der molekularen Marker berechneten Phylogenie wird die Wildrebe aus dem Kaukasus daher als Außengruppe geführt. Zwar werden die kaukasischen Wildreben derzeit der Subspezies *Vitis vinifera* subsp. *sylvestris* zugeordnet, jedoch wird die Einordnung in die eigene Subspezies *Vitis vinifera* subsp. *caucasica* diskutiert (Levadoux 1956). Die hier gezeigten deutlichen Unterschiede bei der Verwendung der Transposon-basierten Marker könnten eine solche Neugliederung der Wildreben möglicherweise unterstützen. Die Ursache für die wenigen Amplifikate in der Marker-PCR wäre dann vermutlich eine Inkompatibilität der Primer, ähnlich wie sie bei der Kartierungsinkompatibilität der *Vitis rotundifolia*-Sequenzdaten beobachtet wurde. Eine andere, simple Erklärung für die stark abweichenden Ergebnisse der Wildrebe aus dem Kaukasus ist eine mögliche Degradation oder eine anders geartete „PCR-Untauglichkeit" der zugrundeliegenden DNA. Bereits die Hochdurchsatzsequenzierung der kaukasischen Wildrebe gestaltete sich schwierig und die Bibliothek generierte deutlich weniger Cluster als erwartet.

4.2.2 Genetische Diversität der Wilden Weinrebe

Die genetische Diversität einer Art beschreibt die Gesamtheit der vererbbaren Variabilität innerhalb und zwischen Populationen dieser Spezies. Sie bildet die Grundlage für die Anpassungsfähigkeit von Organismen an sich verändernde Umweltbedingungen und ist daher für vom Aussterben bedrohte Spezies wie die Wilde Weinrebe von besonderer Bedeutung. Die Wilde Weinrebe stellt zudem einen nahen Verwandten der wirtschaftlich bedeutsamen Kulturreben dar. Die genetische Diversität der Wilden Weinrebe kann demzufolge als Ressource für die Rebenzüchtung betrachtet werden.

Im Rahmen dieser Arbeit wurde die genetische Diversität der Wilden Weinrebe auf Populationsebene analysiert. Zu diesem Zweck wurden Pools aus genomischer DNA von bis zu 15 Individuen mittels Hochdurchsatzverfahren sequenziert. Pool-Sequenzierungen erlauben dabei die Untersuchung ganzer Populationen zu einem Bruchteil des Preises der Sequenzierung einzelner Individuen (Schlötterer et al. 2014). Da die genomische DNA bereits zu Beginn vereint wurde, musste nur eine Sequenzierbibliothek pro Population erstellt werden, teures *Barcoding* der einzelnen Individuen entfiel. Insgesamt wurden vier Populationen aus verschiedenen Habitaten und geographischen Großräumen Europas betrachtet. Um deren Diversität zu analysieren, wurden auf Basis der NGS-Daten genomweit Einzelnukleotidpolymorphismen (*single nucleotide polymorphisms*, SNPs)

identifiziert. Neben der Position der Polymorphismen wurden weiterhin die in der Population vorliegenden Allelfrequenzen ermittelt. Die Parameter des Algorithmus zur Identifizierung der SNPs wurden dabei speziell an die Besonderheiten der Pool-Sequenzierung angepasst. Beispielsweise wurde die Anzahl der zu erwartenden Allele erhöht, da in einer Population signifikant mehr verschiedene Allele auftreten können als in einem einzelnen diploiden Organismus, für den der Algorithmus ursprünglich konzipiert wurde. Die geforderte minimale Allelfrequenz wurde hingegen erniedrigt, um auch Varianten zu identifizieren, die nur in einem einzigen Individuum der Population vorliegen. Dies birgt jedoch Gefahren hinsichtlich der Identifizierung von Falsch-Positiven aufgrund der verhältnismäßig hohen Fehlerrate der Illumina-Sequenzierungstechnologie (Meacham et al. 2011). Die sichere Unterscheidung zwischen Sequenzierfehlern und Polymorphismen mit geringer Allelfrequenz ist daher eine der großen Herausforderungen der Pool-Sequenzierung (Schlötterer et al. 2014). Im Rahmen dieser Arbeit wurden daher verschiedene Strategien verfolgt, um sie zu meistern. Den ersten Schritt stellte hierbei eine strenge Qualitätsfilterung der Illumina-Sequenzen gleich zu Beginn der Analysen dar. Sequenzabschnitte mit schlechter Qualität, die potenziell Sequenzierfehler enthalten können, wurden konsequent aus den Datensätzen entfernt. Zusätzlich erfolgte die Identifizierung der Polymorphismen mit einem sogenannten qualitätsbasierten Algorithmus, der die Phred-Werte der SNP-Positionen sowie der sie umgebenden Basen überprüft. Als letzte Stufe wurde eine Stichprobe von 15 zufällig ausgewählten Polymorphismen experimentell, d. h. mittels Sanger-Sequenzierung, verifiziert. Die Sanger-Sequenzierung wurde weiterhin genutzt, um die bioinformatisch ermittelten Allelfrequenzen in den vorliegenden Populationen zu überprüfen. Idealerweise enthalten die Pools gleiche Mengen an DNA von jedem Individuum. Durch technische Fehler beim Pipettieren oder bei der Quantifizierung der genomischen DNA kann jedoch ein Ungleichgewicht in der Zusammensetzung des Pools entstehen, wodurch die ermittelten Allelfrequenzen verzerrt werden (Schlötterer et al. 2014). Hinzu kommt, dass die verwendeten Pools einen geringen Umfang haben. In einigen Publikationen wurden größere Pools analysiert, wodurch Imbalancen beim Poolen deutlich weniger ins Gewicht fallen (Amaral et al. 2011). Nichtsdestotrotz wurden zahlreiche andere Studien mit ähnlich kleinen oder sogar noch geringeren Individuenzahlen erfolgreich durchgeführt (Montague et al. 2014; Axelsson et al. 2013; Rubin et al. 2010). Im vorliegenden Fall wichen die bioinformatisch ermittelten Allelfrequenzen im Schnitt weniger als 3 % von den realen

Allelfrequenzen der untersuchten Population ab. Dies ist ein durchaus tolerierbarer Wert der sich zudem mit den Literaturangaben deckt (Rellstab et al. 2013).

Zusammengenommen konnten in den vier untersuchten Populationen 23,4 Mio. SNPs identifiziert werden. Je nach Pool ergab sich eine Polymorphismendichte von durchschnittlich einem SNP alle 35 bis 48 bp. Ein Vergleich mit der Literatur ist schwierig, da die vorliegende Arbeit die erste Studie ist, die in diesem Umfang genomweit Polymorphismen in Wildrebenpopulationen identifiziert. Zwar untersuchte eine aktuelle Studie von Marrano et al. (2017) ebenfalls die genetische Diversität einer Population aus 44 *Vitis vinifera* subsp. *sylvestris*-Akzessionen auf Basis von NGS-Daten, jedoch verwendeten die Autoren eine sogenannte *Restriction-site Associated DNA* (RAD)-Sequenzierung, so dass ihnen nur 1,1 % des Rebengenoms für die Analyse zur Verfügung standen. Ältere Studien verglichen einzelne Individuen der Edlen und Wilden Weinreben auf Basis der Sanger-Sequenzierung kurzer genomischer Abschnitte von maximal 100 kb Gesamtlänge. Dabei ergab sich eine SNP-Dichte von einem SNP alle 33 bis 64 bp (Riahi et al. 2013; Lijavetzky et al. 2007; Salmaso et al. 2004). Hierbei ist jedoch zu beachten, dass einige der Autoren ausschließlich kodierende Bereiche analysierten. Wie bereits erwähnt bezogen sich bisherige Analysen von Polymorphismen in der Wilden Weinrebe ausschließlich auf kleine Ausschnitte des Genoms (Marrano et al. 2017) oder es wurden aus Kulturreben bekannte Positionen verwendet (Myles et al. 2011). Mit den über 20 Mio. *de novo* identifizierten SNP-Positionen in der Wilden Weinrebe liegt nun erstmals ein umfassender Katalog der Polymorphismen von *Vitis vinifera* subsp. *sylvestris* vor. Dieser kann nicht nur wie hier beschrieben zur Analyse der genetischen Diversität der Wilden Weinrebe genutzt werden, sondern bietet auch zahlreiche weitere Anwendungsmöglichkeiten. Die identifizierten Positionen könnten sich beispielsweise als hochauflösende molekulare Marker für die Konstruktion von Kopplungskarten, die Identifizierung von QTLs, eine Marker-gestützte Selektion oder die Genotypisierung von Wildreben eignen (Hayward et al. 2015). Gegenüber den im Rahmen dieser Arbeit genutzten SSR- oder Transposon-basierten Markern bieten solche SNP-basierten Marker eine uniformere Verteilung über das gesamte Genom sowie eine höhere Stabilität hinsichtlich Position und Länge als Vorteile (Gupta et al. 2001).

Hinsichtlich der Art der identifizierten Polymorphismen dominieren in den untersuchten Wildreben die Transitions- über den Transversionsmutationen. Dies ist auf den ersten Blick

verwunderlich, da es ausgehend von einer gegebenen Base zwei Möglichkeiten für eine Transversion, jedoch nur eine Möglichkeit für eine Transitionsmutation gibt. Die zwei häufigsten biochemischen Mechanismen, die einen Basenaustausch verursachen, sind Tautomerien und Desaminierung – beides Prozesse, die in einer Transitionsmutation resultieren und so das Verhältnis zugunsten dieses Mutationstyps verschieben (Griffith et al. 2000). Hinzu kommt, dass es sich bei Transitionsmutationen häufiger als bei Transversionsmutationen um synonyme Austausche handelt, die wiederum seltener durch natürliche Selektion entfernt werden (Watson et al. 2008). Das im Rahmen dieser Arbeit ermittelte Ti/Tv-Verhältnis von 1,76 bis 2,03 liegt über den Literaturwerten für die Weinrebe (1,46 bei Lijavetzky et al. 2007; 1,55 bei Salmaso et al. 2004), befindet sich aber durchaus im Rahmen der für Pflanzen beschriebenen Werte (Nickrent & Soltis 1995).

In den untersuchten Wildrebenpopulationen lag ein deutlicher Überschuss von Allelen mit einer minoren Allelfrequenz (MAF) kleiner 0,1 vor. Vezzulli et al. (2008) konnten in den von ihnen untersuchten Wilden Weinreben, eine ähnliche Häufung niedrigfrequenter Allelvarianten beobachten (Abbildung 28). Die parallele Untersuchung der MAF in Kulturreben durch selbige Autoren ergab hingegen eine sichtlich homogenere Verteilung der Allelfrequenzen bei einer durchschnittlichen MAF von 0,3.

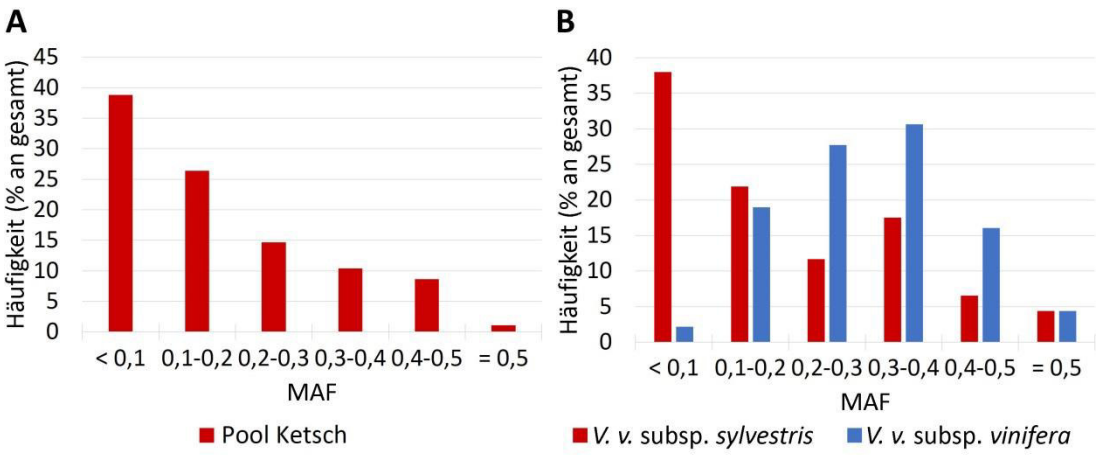

Abbildung 28: Überschuss niedrigfrequenter Allelvarianten in Wilden Weinreben
A zeigt die Häufung von Allelen mit einer MAF kleiner 0,1 unter den im Rahmen dieser Arbeit identifizierten Polymorphismen am Beispiel des Pools Ketsch. **B** zeigt in rot einen ähnlichen Überschuss niedrigfrequenter Allele mit einer MAF kleiner 0,1 in den von Vezzulli et al. (2008) untersuchten Wilden Weinreben. Zum Vergleich sind zusätzlich die von Vezzulli et al. (2008) erhobenen MAFs der Polymorphismen Edler Weinreben in blau aufgetragen.

Der Überschuss sehr seltener Allelvarianten in den untersuchten Pools deutet auf eine geringe Heterozygosität in den zugrundeliegenden Wildrebenpopulationen hin. Die Ursachen für diese genetische Verarmung oder Erosion sind vermutlich in der geringen Individuenzahl der Populationen und der damit verbundenen Inzuchtsproblematik sowie im fehlenden Genfluss zwischen den voneinander isolierten Populationen zu suchen. Letztlich bringt der Verlust der genetischen Diversität die ohnehin bedrohten Wilden Weinreben zusätzlich in Gefahr, da infolgedessen die Anpassungsfähigkeit der Pflanzen an Umweltveränderungen stark eingeschränkt wird (Arroyo-García & Revilla 2013).

Bei einem Vergleich der vier Pools untereinander lässt sich ein Süd-Nord-Gefälle hinsichtlich der Gesamtzahl identifizierter SNPs und einzigartiger Polymorphismen erkennen. Im geographisch am südlichsten gelegenen Pool Spanien wurde mit 13,4 Mio. SNPs nicht nur die größte Gesamtzahl an Polymorphismen identifiziert, sondern auch die höchste Zahl an Pool-spezifischen SNPs. Mehr als 30 % der Polymorphismen stellten eine für den Pool aus Spanien einzigartige Diversität dar. Einen ähnlich hohen Anteil „exklusiver" Allele wurde für den nur geringfügig nördlicher gelegenen Pool Pisa identifiziert. Über höhere Diversität der Wilden Weinreben in Spanien und Italien wird auch in Studien von De Andrés et al. (2012) und Biagini et al. (2014) berichtet. Dieses Phänomen ist als *southern richness* bekannt (Grassi et al. 2008). Deutlich weniger Pool-spezifische Polymorphismen fanden sich hingegen in den untersuchten Wildrebenpopulationen aus Frankreich und Ketsch. Der Pool Frankreich übertraf den Pool aus Ketsch jedoch deutlich in der Gesamtzahl identifizierter SNPs. Die Wilden Weinreben aus Ketsch stellen den nördlichsten Rand des Verbreitungsgebiets von *Vitis vinifera* subsp. *sylvestris* dar. Die geringe genetische Diversität der Ketscher Wildreben deckt sich mit Literaturangaben (Ledesma-Krist et al. 2015). Das hier beobachtete Süd-Nord-Gefälle der genetischen Diversität der Wilden Weinreben ist historisch bedingt: Während der letzten Eiszeit im Pleistozän dienten Regionen im Süden Italiens und der iberischen Halbinsel als Refugien für die Wilde Weinrebe und zahlreiche andere Pflanzenspezies. Im Zuge der nacheiszeitlichen Rekolonialisierung Richtung Norden kam es zu einer sukzessiven Verarmung der genetischen Diversität (Hewitt 1999). Pflanzen, die sich auf die Gegebenheiten im Norden, wie beispielsweise kältere Temperaturen, besser einstellen konnten, hatten einen Selektionsvorteil gegenüber den nur an südliche Gefilde angepassten Reben, wodurch es zu einem sogenannten Flaschenhalseffekt kam (Ledesma-Krist et al. 2015). Neben dem Süd-Nord-Gefälle lässt sich auch eine putative

Route der Rekolonialisierung aus den vorliegenden Daten der vier Wildrebenpopulationen ableiten. Die größte Gemeinsamkeit in Form von Pool-übergreifenden Polymorphismen zeigen die beiden Pools aus Spanien und Frankreich. Daher fand die Rekolonialisierung Frankreichs vermutlich ausgehend von den Refugien im Süden der iberischen Halbinsel statt. Für die nördlicheren Wildrebenpopulationen schlagen Grassi et al. (2008) eine Route ausgehend von Italien vor. Dem widersprechen die vorliegenden Daten, da die Pools Ketsch und Pisa die geringste Übereinstimmung durch Pool-übergreifende SNPs aufweisen. Die Gemeinsamkeiten von Ketsch und Frankreich sind hingegen größer, wodurch die Fortsetzung der in Spanien entspringenden Rekolonialisierungsroute über Frankreich hinaus nach Deutschland wahrscheinlicher wird. Nicht auszuschließen ist jedoch auch ein Ursprung in östlichen Refugien wie beispielsweise im Balkan oder dem Kaukasus, für die im Rahmen dieser Arbeit keine Polymorphismus-Daten erhoben wurden.

4.3 Die Wilde Weinrebe als genetische Ressource

Die Wilde Weinrebe *Vitis vinifera* subsp. *sylvestris* ist ein naher Verwandter der Edlen Weinrebe *Vitis vinifera* subsp. *vinifera*, der Kulturrebe, die rund um die Welt als einzige Spezies der Gattung *Vitis* für die Weinproduktion angebaut wird (This et al. 2006). Die beiden Subspezies sind fertil kreuzbar (Aradhya et al. 2008). *Vitis vinifera* subsp. *sylvestris* gehört somit gemäß dem Genpool-Konzept von Harlan und De Wet (1971) zum GP-1b der Kulturrebe und kann daher als genetische Ressource für die Rebenzüchtung genutzt werden. Bei den Merkmalen der Wilden Weinrebe, die für Winzer und Züchter von Interesse sind, handelt es sich zumeist um Resistenzeigenschaften in Bezug auf Pilzinfektionen und Viruserkrankungen (Nick 2012; Arnold et al. 1998). Schröder et al. (2015) identifizierten unter den Wilden Weinreben der Halbinsel Ketsch einzelne Individuen mit einem besseren Resistenzverhalten gegenüber dem Echten Mehltau (*Erysiphe necator*) und dem Falschen Mehltau (*Plasmopara viticola*) als die Kulturrebe Regent. Aber auch eine Bedeutung bei der Verbesserung der Thermotoleranz wird für *Vitis vinifera* subsp. *sylvestris* diskutiert.

Ein unkontrolliertes Verkreuzen Wilder und Edler Weinreben ähnlich einer Spontanbastardisierung ergibt aufgrund der zahlreichen negativen Eigenschaften von *Vitis vinifera* subsp. *sylvestris* wenig Sinn (Maul 2016). Um die Wilde Weinrebe sinnvoll als genetische Ressource nutzen zu können, sollte man die der gewünschten Eigenschaft

zugrundeliegende genomische Region und bestenfalls die kausative Variante identifizieren. Bei bekannten Resistenzmechanismen, wie beispielsweise der Akkumulation antimikrobieller Stilbene, kann dies verhältnismäßig einfach sein. In so einem Fall bietet es sich an, die homologen Gene in der Wilden Weinrebe mittels degenerierter Primer oder BLAST-Suchen gegen *de novo* Assemblierungen zu identifizieren und mit den Kulturreben-Varianten zu vergleichen. Auf diese Art und Weise konnte mit der im Rahmen dieser Arbeit generierten *de novo* Assemblierung der Illumina-Sequenzen von Hördt29 ein Resistenzmechanismus der Wilden Weinrebe gegen den Falschen Mehltau (*Plasmopara viticola*) aufgeklärt werden (Duan, Fischer et al. 2016). Unbekannte Mechanismen und Gene lassen sich hingegen auf diesem Weg nicht ausfindig machen. Daher wurde hierfür die Strategie gewählt, nach Fußspuren der Selektion im Genom der Wilden Weinrebe zu suchen. Da die Wilde Weinrebe gleichermaßen dem Selektionsdruck durch die natürlich vorkommenden Pathogene ausgesetzt ist, aber im Gegensatz zur Kulturrebe nicht mit Pflanzenschutzmitteln behandelt wird und sich zudem noch sexuell fortpflanzt, sollte es möglich sein, auf diese Art und Weise interessante Kandidatengene für die Rebenzüchtung zu identifizieren.

4.3.1 Selektionsmuster im Genom der Wilden Weinrebe

Selektion hinterlässt molekulare Fußspuren im Genom (Oleksyk et al. 2010). Wenn eine vorteilhafte Mutation selektiert wird, steigt ihre Frequenz in der Population und mit ihr die Frequenz gekoppelter neutraler Allele (Maynard Smith & Haigh 1974). Die Variationsstruktur der betroffenen genomischen Region verändert sich und ein sogenannter *selective sweep* entsteht (Sabeti et al. 2006). Die Veränderungen in der Variationsstruktur können somit genutzt werden, um Selektionsereignisse im Genom zu identifizieren. Dies wurde im Rahmen dieser Arbeit durch Analyse der gepoolten Heterozygosität (H_p) für das Genom der Wilden Weinrebe realisiert. Diese von Rubin et al. (2010) erstmals vorgeschlagene Strategie identifiziert *selective sweeps* anhand der lokalen Reduktion der Heterozygosität des entsprechenden DNA-Abschnitts. Die Reduktion wird gleichermaßen durch die Selektion vorteilhafter Mutationen und ihrer gekoppelten Varianten sowie durch den Überschuss seltener Varianten infolge der langsamen Rückkehr der Diversität nach der Fixierung der selektierten Mutation in der Population verursacht. Die Erhebung der Daten erfolgte mittels eines Fensters, das an der Sequenz des Genoms entlanggleitet und bei jedem Halt die gepoolte Heterozygosität aller Positionen innerhalb

des Fensters berechnet. Der Vorteil des Fensters ist also, dass die Information zahlreicher benachbarter Positionen kombiniert und infolgedessen das Hintergrundrauschen der einzelnen Loci reduziert wird (Qanbari et al. 2012).

In einigen Studien wurde die Größe des Fensters ohne Vorwissen auf Basis von Erfahrungswerten anderer Autoren, die jedoch häufig anderere Spezies untersuchten, festgelegt (Choi et al. 2015). Eine falsche Wahl kann jedoch die Ergebnisse stark beeinflussen und verzerren. Genetisch betrachtet hängt die Größe eines *selective sweeps* u. a. vom Umfang des Kopplungsungleichgewichts (*linkage disequilibrium,* LD) ab (Qanbari et al. 2012). Korrekterweise sollte die Fenstergröße daher die Ausdehnung des Kopplungsungleichgewichts in der untersuchten Spezies reflektieren (Stölting et al. 2015). Die kurzen *read*-Längen der vorliegenden Illumina-Sequenzdaten eignen sich jedoch nicht für die Abschätzung des LDs in der Wilden Weinrebe (Schlötterer et al. 2014). Daher wurde an dieser Stelle auf Literaturangaben zurückgegriffen. Marrano et al. (2017) ermittelten in der von ihnen untersuchten Wildrebenpopulation einen LD von 20 kb, Nicolas et al. (2016) geben hingegen Werte von 31 bis 127 kb an. Neben dem Kopplungsungleichgewicht mussten noch weitere Faktoren bei der Wahl der Fenstergröße berücksichtigt werden. Bei dem Referenzgenom, das für die Kartierung verwendet wurde, handelt es sich um einen sogenannten Genom-Entwurf, der aus vielen einzelnen *contigs* besteht. Diese sind zwar in Gerüststrukturen, die Chromosomen imitieren, zusammengefasst, jedoch *in vivo* durch Lücken unbekannter Größe getrennt. Die verwendeten Fenster sollten auf keinen Fall die genomischen Abschnitte 5‘ und 3‘ einer Lücke zusammenfassen, da in diesem Fall Bereiche gemeinsam analysiert werden würden, die unter Umständen gar nicht gekoppelt sind. Daher wurden für die Berechnung der gepoolten Heterozygosität die *contigs* aus der Gerüststruktur herausgelöst und einzeln untersucht. Jedoch ist in diesem Fall eine Analyse nur möglich, wenn die Länge des *contigs* die gewählte Fenstergröße übersteigt. Kleinere *contigs* sind von den Untersuchungen ausgeschlossen. Zu kleine Fenster beinhalten jedoch nur eine geringe Anzahl polymorpher Loci und können daher nicht die tatsächliche Heterozygosität der zu analysierenden Region repräsentieren (Sun et al. 2014). Unter Berücksichtigung all dieser Faktoren wurde die Fenstergröße für die Analyse der gepoolten Heterozygosität im Genom der Wilden Weinrebe auf 40 kb festgelegt. Damit lag sie zwar über dem beobachteten Kopplungsungleichgewicht von Marrano et al. (2017), befand sich aber im Bereich der Werte von Nicolas et al. (2016). Mit im Schnitt 800 bis 1.200 SNPs

waren ausreichend polymorphe Positionen vorhanden und der Verlust kleiner *contigs* befand sich in einem vertretbaren Rahmen.

Die identifizierten Kandidatenregionen zeigten im genomweiten Diversitätsdiagramm ein charakteristisches Aussehen. Es handelte sich zumeist um Cluster aus mehreren benachbarten Fensten mit erniedrigten H_P-Werten, wobei die H_P-Werte entlang der Fenster zunächst schrittweise abnahmen, ein Minimum erreichten um anschließend wieder schrittweise anzusteigen. Dieser Tal-förmige Verlauf ist typisch für *selective sweeps*. Im Zentrum des *sweeps* ist die Reduktion der genetischen Diversität aufgrund der dort angreifenden Selektionskraft am stärksten. Mit steigender Entfernung nimmt die Rekombination zu und die Diversität steigt an (Kim & Stephan 2002). Die beobachteten Muster unterschieden sich in den einzelnen Pools voneinander. Dies ist darin begründet, dass die verschiedenen Populationen in ihren jeweiligen Habitaten spezifischen Selektionsdrücken ausgesetzt sind. Die Ausprägung eines *selective sweeps* wird zudem durch Faktoren beeinflusst, die sich in den Populationen unterscheiden können. Dazu gehören das Alter des Selektionsereignisses, lokale Unterschiede in der Rekombinationsrate, der erreichte Fixierungsgrad der selektierten Variante und die Populationsstruktur zu Beginn der Fixierung (Qanbari et al. 2012). Zwar wurde beim experimentellen Design darauf geachtet, die Poolgröße mit 10 bis 15 Individuen möglichst vergleichbar zu gestalten, jedoch unterscheiden sich die zugrundeliegenden Populationen hinsichtlich Größe und geographischer Ausbreitung voneinander. Die beiden Pools Ketsch und Pisa stammen z. B. aus relativ kleinen isolierten Habitaten, die Pools Frankreich und Spanien hingegen aus größeren geographischen Räumen. Dies könnte einen Einfluss auf das beobachtete Diversitätsmuster bei der Pool-Sequenzierung gehabt haben.

Zur Identifizierung der *selective sweeps* wurde in zahlreichen Studien eine sogenannte *outlier*-Strategie angewendet, bei der die Fenster, die sich in den extremen Minima der ZH_P-Verteilung befinden, als relevant betrachtet werden (Voight et al. 2006). Jedoch werden bei dieser Herangehensweise testbare Hypothesen vermieden und die statistische Signifikanz der identifizierten Selektionssignale kann nicht bewertet werden. Daher wurde im Rahmen dieser Arbeit auf die *outlier*-Strategie verzichtet und anstelle dessen eine Methode von Qanbari et al. (2012) angewendet. Diese überprüft, ob ein Signal auch durch Zufall, also in Abwesenheit von Selektion, entstehen kann. Zu diesem Zweck erfolgte eine Simulation des Zufalls indem die erhobenen Polymorphismen des jeweiligen Datensatzes permutiert und

daraus eine Nullverteilung abgeleitet wurde. Anhand dieser konnte die genomweite Signifikanz getestet und ein Schwellenwert berechnet werden. Der Vorteil der Methode ist, dass bei der Permutation das originale Frequenzspektrum und die genomische Struktur in Form der SNP-Positionen beibehalten wird (Qanbari et al. 2012). So konnte für jeden Datensatz ein spezifischer Schwellenwert ermittelt werden, der die jeweilige Populationsstruktur berücksichtigt. Insgesamt wurden in den einzelnen Pools zwischen 466 und 1.416 Fenster mit statistisch signifikant reduzierten H_p-Werten identifiziert ($p \leq 0{,}001$). Ein Großteil dieser Fenster wäre bei Anwendung der *outlier*-Strategie übersehen worden, bei der typischerweise nur die „Top 1 %"-Fenster mit den kleinsten H_P-Werten als relevant gewerten werden – im vorliegenden Fall wären das in den einzelnen Pools 363 bis 369 Fenster gewesen.

Im Pool Spanien wurde die größte Zahl putativer Selektionsereignisse identifiziert. Dieser Pool stach bereits bei der Analyse der genetischen Diversität durch eine große SNP-Gesamtzahl und einen großen Anteil Pool-spezifischer Polymorphismen hervor. Daher kann spekuliert werden, dass die in der Population vorliegende stehende Variation mehr Angriffspunkte für die Selektion bot und sie sich daher besser an die gegebenen Umweltbedingungen anpassen konnte. Dieser Zusammenhang bestätigt sich auch in den anderen Pools. Die wenigsten Fußspuren der Selektion wurden in den Wildreben der Halbinsel Ketsch nachgewiesen. Diese Population zeigt bereits eine geringe genetische Diversität. Nur die beiden verbleibenden Pools haben die Plätze getauscht. Im Pool Frankreich wurden geringfügig mehr putative Selektionssignale als im Pool Pisa identifiziert, letzterer zeigte aber die größere Vielfalt bei der Analyse der Polymorphismen. Dies könnte im Zusammenhang mit der bereits diskutierten nacheiszeitlichen Rekolonialisierungsroute von Spanien nach Frankreich stehen. Auf dem Weg Richtung Norden ging durch Flaschenhalseffekte zwar ein Teil der genetischen Diversität verloren, bereits in Spanien vorhandende Angriffspunkte der Selektion blieben jedoch aufgrund des anhaltenden Selektionsdrucks erhalten.

Die im Rahmen dieser Arbeit verwendete Strategie zur Identifizierung genomweiter Fußspuren der Selektion zeigt hinsichtlich der im folgenden beschriebenen Aspekte einige Einschränkungen. Ausgangspunkt der Analysen stellte das Referenzgenom der Kulturrebe Pinot Noir dar, das für die Kartierung der Wildrebensequenzen verwendet wurde. Sequenzdaten von Wildreben-exklusiven Genombereichen oder DNA-Abschnitten in denen

sich Kultur- und Wildreben signifikant unterscheiden konnten somit nicht im Referenzgenom kartieren. Dabei handelte es sich nach Abzug der Chloroplasten- und Mitochondrien-Sequenzen um etwa 12 bis 15 % der Gesamtdaten der Wilden Weinreben. Eine Lösung hierfür wäre gewesen, anstelle des Referenzgenoms der Kulturrebe eine *de novo* Assemblierung einer Wilden Weinrebe, beispielsweise die des Hördt29, als Referenz für die Kartierung zu verwenden. Jedoch war die durchschnittliche *contig*-Länge dieser Assemblierung so gering, dass eine Anpassung der Fenstergröße an die Ausdehnung des Kopplungsungleichgewichts der Wilden Weinrebe unmöglich gewesen wäre. Darüber hinaus stand für diese Assemblierung keine Genannotation zur Verfügung, was die weitere Auswertung erschwert hätte. Ferner war die verwendete Strategie klar auf die Selektion kleiner Veränderungen wie SNPs oder kurze Indels ausgerichtet. Größere strukturelle Veränderungen im Genom lassen sich aufgrund der kurzen *read*-Längen der Illumina-Sequenzierung nur schwer erkennen. Auch genomische Abschnitte mit repetitivem Charakter, wie beispielsweise transposable Elemente, waren von der Analyse ausgeschlossen, da mehrfach kartierbare *reads* oder Genomabschnitte mit extrem hoher Abdeckung nicht für die Identifizierung der Polymorphismen verwendet wurden.

Einige Studien nutzen bekannte *quantitative trait loci* (QTLs) zum Nachweis der grundsätzlichen Machbarkeit (*proof of priciple*) der Selektionsanalysen (Rubin et al. 2010; Gheyas et al. 2014). In der Wilden Weinrebe existieren jedoch kaum bekannte QTLs. Die typischen Resistenz-vermittelnden QTLs stammen aus amerikanischen und asiatischen *Vitis*-Spezies (Schwander et al. 2012; Donald et al. 2002). Daher wurde an dieser Stelle auf einen solchen Nachweis verzichtet. Jedoch war aus einer Kooperation mit dem Botanischen Institut des Karlsruher Instituts für Technologie bekannt, dass in einigen Ketscher Wildreben Stilbensynthasen und ihre Regulatoren eine Resistenz gegenüber dem Falschen Mehltau (*Plasmopara viticola*) vermitteln (Duan et al. 2016). Die entsprechenden Gene befanden sich jedoch nicht in den identifizierten Kandidatenregionen des Pools Ketsch. Jedoch handelt es sich bei den Genen der Stilbensynthasen und ihrer Regulatoren, den MYB-Genen, um Mitglieder großer Multigenfamilien, was eine spezifische Kartierung und Identifizierung von Polymorphismen erschwert (Fischer 2012).

Die vorliegende Studie ist die erste Analyse von Selektionssignalen im Gesamtgenom der Wilden Weinrebe auf Basis von NGS-Daten. Insgesamt konnten in den vier untersuchten Populationen zwischen 153 und 434 genomische Kandidatenregionen identifiziert werden,

die vermutlich unter Selektion standen. Eine nötige Absicherung dieser Signale erfolgt am besten durch weitere Selektionsanalysen. Im Huhn, die erste Spezies für die eine Untersuchung der gepoolten Heterozygosität nach Rubin et al. (2010) durchgeführt wurde, wurden bis heute mehr als fünf Selektionsanalysen publiziert, von denen einige wiederum die identifizierten Signale ihrer Vorgänger bestätigten (z. B. Guo et al. 2016). Dabei empfielt sich die kombinierte Anwendung verschiedener Detektionsmethoden jenseits der gepoolten Heterozygosität, wie beispielsweise dem Fixierungsindex F_{ST} oder Tajima's D (Carneiro et al. 2014; Johnsson et al. 2016). Lassen sich mit verschiedenen Herangehensweisen wiederholt Signale feststellen, liefert dies stichhaltigere Hinweise auf ein tatsächlich stattgefundenes Selektionsereignis. Im Fall der Wilden Weinrebe sollten weitere Studien darüber hinaus auch Populationen aus dem Balkan und dem Kaukasus umfassen, da diese laut Literatur eine besonders hohe genetische Diversität beinhalten und geographisch betrachtet anderen Selektionsdrücken ausgesetzt waren (De Mattia et al. 2008).

4.3.2 Kandidatengene für die Anwendung in der Rebenzüchtung

In den anhand der molekularen Fußspuren der Selektion identifizierten Kandidatenregionen sind in allen vier Pools zusammengenommen 2.601 Gene annotiert. Diese Annotation beruht auf dem Referenzgenom der Kulturrebe Pinot Noir und umfasst ausschließlich proteinkodierende Gene. Gene für miRNAs und andere nicht-kodierende RNAs sind nicht enthalten. Ein Teil der Kandidatenregionen beinhaltet keine Gene, jedoch ist nicht auszuschließen, dass dies auf eine unvollständige Annotation des Referenzgenoms zurückzuführen ist. Für eine hier nicht untersuchte genomische Region konnte eine solche fehlende Annotation bereits gezeigt werden (Ding 2016). Eine manuelle Annotation, beispielsweise anhand von BLAST-Suchen oder mit Hilfe von Algorithmen zur Identifizierung von offenen Leserahmen, wäre möglich, wurde hier jedoch aufgrund des großen Umfangs der Daten nicht vorgenommen. Der Gengehalt der identifizierten Kandidatenregionen weicht in zwei Pools geringfügig von dem des Gesamtgenoms ab, was nicht ungewöhnlich ist, da Selektion auch jenseits von Genen z. B. in regulatorischen Sequenzen wirken kann (Moses 2009). In einem Großteil der identifizierten Kandidatenregionen sind mehrere Gene annotiert. Es ist auszuschließen, dass all diese Gene einer Selektion unterlagen. Vielmehr ist dies auf die Fenstergröße von 40 kb und den sogenannten *hitchhiking effect* zurückzuführen. Aufgrund der genetischen Kopplung werden bei der Selektion einer

vorteilhaften Variante neutrale Varianten in der direkten Umgebung mitselektiert (Maynard Smith & Haigh 1974). Befinden sich diese neutralen Varianten in benachbarten Genen erstreckt sich das Selektionssignal über diese hinweg. Eine Entscheidung, welches der jeweiligen Gene im Zentrum des *selective sweeps* liegt, ist ohne weiterführende Analysen der genomischen Regionen nicht möglich. Hierfür wäre beispielsweise ein Scan mit kleineren Fenstergrößen oder die Verwendung anderer Detektionsmethoden, wie die Betrachtung des Verhältnisses von nicht-synonymen zu synonymen Substitutionen innerhalb der Gene, notwendig (Rubin et al. 2010; Nei 2005).

Die große Anzahl identifizierter Gene erforderte weitere Auswertestrategien, die den Umfang der Zahl der Gene reduzierten und bestenfalls relevante Gene herausfilterten. Zugleich wurden die Gene funktionellen Gruppen (*Gene Ontology-Terms*) zugeordnet, die Rückschlüsse auf ihre molekulare Funktion und ihre Beteiligung an biologischen Prozessen erlauben. Im Rahmen einer der vier Auswertestrategien wurden funktionelle Gruppen identifiziert, die unter den analysierten Genen im Vergleich zum Gesamtgenom statistisch angereichert und somit überrepräsentiert waren. Drei Gruppen stechen dabei besonders hervor, von denen die Gengruppe, die mit dem Transport von Kupferionen assoziiert ist, am ehesten eine Relevanz für die Rebenzüchtung besitzt. Auf Kupfer basierende Pflanzenschutzmittel werden seit Ende des 19. Jahrhunderts im Weinbau als Fungizid verwendet (Kühne et al. 2009). In den vergangenen zwanzig Jahren ist eine Zunahme der eingesetzten Kupfermengen durch die Ausweitung des ökologischen Weinbaus zu verzeichnen (Berkelmann-Löhnertz et al. 2008). Aufgrund des erhöhten Kupfereintrags in den Boden könnte eine Züchtung kupfertoleranter Weinreben von Interesse sein. Jedoch soll langfristig aufgrund assoziierter ökotoxikologischer Aspekte auch im Bio-Weinbau auf Kupferpräparate verzichtet werden (Diesner et al. 2014). Daher ist die Nachfrage nach einer solche Züchtungsform der Weinrebe vermutlich gering.

Eine zweite Auswertestrategie analysierte die Gene der Kandidatenregionen, die sich in den extremen Minima der ZH_P-Verteilung befanden. Diese Kandidatenregionen zeigten demzufolge genomweit den höchsten Grad der Fixierung (Rubin et al. 2010). Unter den 54 Genen dieser Regionen befinden sich zahlreiche Haushalts- und Strukturgene wie Gene für ribosomale Proteine oder Bestandteile des Lichtsammelkomplexes der Chloroplasten. Die Produkte dieser Gene sind essenziell für das Überleben der Pflanze, neue entstehende nachteilige Varianten werden unverzüglich durch reinigende Selektion aus der Population

entfernt, was die Reduktion der Diversität am entsprechenden Locus erklärt (Charlesworth et al. 1993). Jedoch wurden mit Hilfe dieser Strategie auch sieben Gene identifiziert, deren zugeordneten *GO-Terms* eine mögliche Relevanz für die Rebenzüchtung andeuten. Bei beiden Genen mit Bezug zu Abwehrmechanismen (GSVIVG01009369001, ZH_P=-2,94 und GSVIVG01023432001, ZH_P=-3,84) ergaben Literaturrecherchen zu den Homologen aus *Arabidopsis thaliana* jedoch eine relativ unspezifische Beteiligung an generellen Stress- und Pathogen-induzierten Abwehrreaktionen (Zheng et al. 2012; Meyers 2003). Ein Zusammenhang mit Resistenzmechanismen gegenüber typischen Weinrebenpathogenen ließ sich nicht herstellen. Keines der drei Kandidatengene, die aufgrund ihrer möglichen Rolle bei Fortpflanzungsprozessen identifiziert wurden, befindet sich im Bereich der geschlechtsbestimmenden Region der Weinrebe auf Chromosom 2. Unter den Genen, die Anpassungsreaktionen auf sich verändernde Umweltfaktoren vermitteln, fiel besonders das Gen GSVIVG01003479001 (ZH_P=-3,50) auf, da dessen *Arabidopsis thaliana*-Homolog die Expression von *cold-regulated* (COR)-Genen während der Akklimatisation an kältere Temperaturen aktiviert (Pavangadkar et al. 2010). Ein solches Gen hätte für die Züchtung kältetoleranter Weinreben, die sich für einen Anbau in nördlichen Regionen eignen, Potenzial. Durch eine Überexpression von *AtDREB1*, einem anderen Regulator der COR-Gene von *Arabidopsis thaliana*, konnte bereits eine Kälteresistenz in transgenen Weinreben vermittelt werden (Jin et al. 2009).

Weiterhin wurden im Rahmen der dritten Auswertestrategie diejenigen Gene betrachtet, die in allen vier untersuchten Pools auftraten und dadurch eine Art Knotenpunkt darstellen. Die Genzahl wurde bei dieser Vorgehensweise im Vergleich zu den anderen Strategien am deutlichsten reduziert, da im Anschluss nur noch 33 „Kerngene“ verblieben. Darunter befand sich jedoch kein Gen mit Beteiligung an pflanzlichen Abwehrmechanismen gegenüber Pathogenen. Erneut war keines der identifizierten Gene mit Bezug zu Fortpflanzungsprozessen auf Chromosom 2 lokalisiert. Das einzige Gen unter den Kerngenen, was mit der Vermittlung einer Thermotoleranz in Verbindung steht, wurde bereits im Zuge der Auswertung der extremen Minima identifiziert. Dies lässt schließen, dass eine Reduktion der identifizierten Kandidatengene auf die „Kerngene“ zu stringent für eine Identifizierung Züchtungs-relevanter Gene war. Jedoch lässt das Vorhandensein von 33 „Kerngenen“ den Schluss zu, dass sich alle vier untersuchten Pools genomische Abschnitte teilen, in denen Signaturen von Selektionsereignissen nachgewiesen wurden. Dies wirft die Frage auf, ob diese das Resultat einer gemeinsamen Selektionsgeschichte oder einer

konvergenten Entwicklung sind. Um dies zu klären, wurde in drei „Kerngenen“ untersucht, ob in den Wilden Weinreben im Vergleich zur Kulturrebe Pinot Noir fixierte Varianten vorliegen. Für zwei der drei Gene konnten insgesamt sieben in den Wildreben fixierte Polymorphismen nachgewiesen werden, darunter zwei nicht-synonyme Nukleotidaustausche. Daher kann für zumindest diese beiden „Kerngene“ eine gemeinsame Selektionsgeschichte der vier untersuchten Populationen vermutet werden.

Zuletzt wurden aus den 2.601 identifizierten Genen all diejenigen Gene ausgewählt, die funktionellen Gruppen mit besonderer Relevanz für die Züchtung angehörten. Darunter befanden sich Gene, die Abwehrmechanismen gegenüber Pathogenen und Krankheiten vermitteln sowie Gene, die eine Rolle bei Fortpflanzungsprozessen spielen und letztlich auch Gene, die für Anpassungsreaktionen auf sich verändernde Umweltfaktoren wie Temperatur oder Wasserverfügbarkeit verantwortlich sind. Diese Vorgehensweise reduzierte zwar die Genzahl in deutlich geringerem Maß als die vorausgegangenen drei Auswertestrategien, sie lieferte aber so die umfassendste Kollektion von Kandidatengenen der Wilden Weinrebe mit züchterischer Relevanz. Zur genaueren funktionellen Charakterisierung dieser Gene wurden, falls möglich, Homologe in *Arabidopsis thaliana* identifiziert und Literaturrecherchen durchgeführt. Als vielversprechendes Kandidatengen kristallisierte sich das Gen GSVIVG01018223001 heraus, das für eine E3 Ubiquitin-Protein Ligase namens BRE1-like 1 kodiert. Es wurde, wie in Abbildung 29 gezeigt, anhand von zwei Selektionssignalen mit signifikant reduzierter genetischer Diversität (ZH_p=-2,45 bzw. -2,63) in den Pools Ketsch und Pisa auf Chromosom 15 identifiziert.

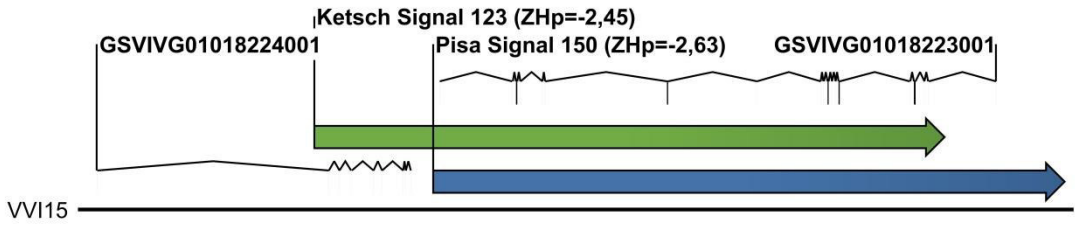

Abbildung 29: Lage des Kandidatengens GSVIVG01018223001 innerhalb der Selektionssignale der Pools Pisa und Ketsch
Gezeigt ist die Position der beiden überlappenden Selektionssignale der Pools Ketsch und Pisa auf Chromosom 15. Im genomischen Abschnitt, der von beiden Selektionssignalen abgedeckt wird, ist das Gen GSVIVG01018223001 annotiert. Darüber hinaus befindet sich im 5‘-Bereich des Ketsch-Signals ein weiteres Gen namens GSVIVG01018224001. Aus Übersichtsgründen ist nur die CDS der Gene abgebildet.

Die beiden Selektionssignale überlappen auf einer Strecke von 32 kb. Alle Exons mit Ausnahme des Exons 1 des entsprechenden Gens befinden sich in dem Sequenzabschnitt, der durch beide Selektionssignale abgedeckt wird. Das abgeleitete Proteinprodukt besteht aus 879 Aminosäuren und enthält eine für BRE1-Proteine typische RING-Finger Domäne (Hwang et al. 2003). Eine BLASTp-Suche identifizierte Homologe in zahlreichen anderen sequenzierten Pflanzengenomen, darunter Pappel, Baumwolle, Sojabohne und *Arabidopsis thaliana*. Das Gen und sein Proteinprodukt HUB1 wurden von Dhawan et al. (2009) in *Arabidopsis thaliana* funktionell charakterisiert. *Loss-of-function* Mutanten *HUB1* sind extrem anfällig für *Botrytis cinerea*, was sich durch verstärkte Chlorosen und Gewebeschäden innerhalb von drei Tagen nach der Infektion manifestiert. Eine Überexpression von *HUB1* äußert sich hingegen in einer Resistenz gegenüber *Botrytis cinerea*. Der Grauschimmelpilz *Botrytis cinerea* befällt zahlreiche Pflanzenspezies rund um die Welt, in der Weinrebe ist er durch die von ihm verursachte Rohfäule der unreifen Beeren für große wirtschaftliche Verluste verantwortlich (Elad et al. 2007). Daher stellt der hier identifizierte Locus der Wildreben aus Ketsch und Pisa eine vielversprechende genetische Ressource für die Rebenzüchtung dar.

5 Zusammenfassung

Weinreben werden weltweit auf einer Fläche von rund 7,5 Mio. Hektar angebaut von der jährlich über 75 Mio. Tonnen Trauben geerntet und in erster Linie zu Wein weiterverarbeitet werden. Sich verändernde Umwelteinflüsse wie Klimawandel und neue Pathogene stellen dabei eine große Herausforderung dar. Um ihr zu begegnen, kann sich die Rebenzüchtung neuer genetischer Ressourcen wie der Wilden Weinrebe *Vitis vinifera* subsp. *sylvestris* bedienen. Im Rahmen dieser Arbeit sollte mithilfe umfassender komparativer Genomanalysen das Potenzial der Wilden Weinrebe beurteilt werden. Zu diesem Zweck wurden mittels *Next-Generation Sequencing* rund 140 Gb Sequenzinformation von 55 Wilden Weinreben, einer Kulturrebe sowie einer Hybridrebe generiert. Dies entspricht einer knapp 300-fachen Abdeckung des Weinrebengenoms. Mittels Kartierungs- und Assemblierungsstrategien wurden die erhaltenen Daten in einen größeren genomischen Kontext gebracht. Die *de novo* Assemblierung des Wildrebengenoms erwies sich jedoch aufgrund der starken Heterozygotie und des hohen Anteils repetitiver Elemente als nicht realisierbar und wurde daher nicht weiter verfolgt. Die Kartierung gegen vorhandene genomische Referenzen der Kulturrebe gestaltete sich erfolgreicher und im Schnitt konnten 95 % des Referenzgenoms mit Sequenzen der Wilden Weinrebe abgedeckt werden.

Transposable Elemente stellen einen hochvariablen Teil des Genoms dar und spielen eine wichtige Rolle in der Architektur und Evolution von Pflanzengenomen. Im Rahmen der vorliegenden Arbeit wurde daher eine Strategie entwickelt, die es ermöglicht, mithilfe von Kartierungen der NGS-Daten die absoluten Kopienzahlen der transposablen Elemente im Genom zu ermitteln. Dabei wurde gezeigt, dass sich die verschiedenen Klassen und Familien transposabler Elemente hinsichtlich der Kopienzahl voneinander und zwischen den Rebengenomen unterscheiden. Aufgrund dieser Unterschiede eignen sich transposable Elemente als molekulare Marker. Daher wurde in der vorliegenden Arbeit der Nutzen ebendieser molekularen Marker zur Unterscheidung der beiden Subspezies *Vitis vinifera* subsp. *vinifera* und *Vitis vinifera* subsp. *sylvestris* getestet und molekulare Fingerabdrücke verschiedener Weinreben beider Subspezies erstellt.

Infolge der anhaltenden, massiven Zerstörung und Fragmentierung ihrer natürlichen Habitate ist die Wilde Weinrebe europaweit vom Aussterben bedroht. Welchen Einfluss

dieser Zustand auf die genetische Diversität der verbleibenden Wildrebenpopulationen hat, wurde im Rahmen dieser Arbeit anhand von vier Populationen aus verschiedenen Habitaten und geographischen Großräumen Europas untersucht. Dabei wurden über 20 Mio. SNP-Positionen *de novo* identifiziert und so erstmals ein genomweiter Überblick über Einzelnukleotidpolymorphismen in der Wilden Weinrebe geschaffen. Beim Vergleich der Pools untereinander zeigt sich ein Süd-Nord-Gefälle hinsichtlich der genetischen Diversität. Dieses Phänomen ist aus der Literatur als *southern richness* bekannt und hat seinen Ursprung in der Verbreitungsgeschichte der Wilden Weinrebe, da die Habitate in Italien und Spanien während der letzten Eiszeit als Refugien dienten, von denen aus die mit Flaschenhalseffekten verbundene Rekolonialisierung Richtung Norden stattfand. Hinsichtlich des geographischen Verlaufs dieser Rekolonialisierung widersprechen die erhobenen Daten jedoch der Literatur und favorisieren eine in Spanien und nicht in Italien entspringende, über Frankreich hinwegreichende Route nach Deutschland. Darüber hinaus konnte in allen vier Populationen übereinstimmend ein Überschuss seltener Allelvarianten mit minoren Allelfrequenzen < 0,1 festgestellt werden. Dieser Befund kann als Indiz einer genetischen Erosion der analysierten Wildrebenpopulationen gedeutet werden – vermutlich infolge der Inzuchtsproblematik innerhalb der kleinen Populationen sowie des fehlenden Genflusses zwischen den voneinander isolierten Populationen.

Für die Rebenzüchtung von besonderem Interesse sind Gene, die Resistenzen gegenüber Pathogenen verleihen, mit der Geschlechtsbestimmung im Zusammenhang stehen oder Reaktionen auf sich verändernde Umweltbedingungen vermitteln können. Um solche Gene in der Wilden Weinrebe zu identifizieren, wurde ihr Genom im Rahmen dieser Arbeit nach molekularen Fußspuren der Selektion durchsucht. In den vier Pools wurden dabei zwischen 153 und 434 mögliche *selective sweeps* identifiziert, die insgesamt 2.601 verschiedene proteinkodierende Gene enthalten. Es wurden verschiedene Strategien vorgestellt, um unter diesen 2.601 Genen geeignete Kandidatengene für die Züchtung zu ermitteln. Der vielversprechendste Kandidat ist das Gen für eine E3 Ubiquitin-Protein Ligase namens BRE1-like 1. In *Arabidopsis thaliana* vermittelt die Überexpression des homologen Gens eine Resistenz gegenüber *Botrytis cinerea*, einem Pathogen vieler Pflanzenspezies, darunter auch die Weinrebe. Für eine Bestätigung des Zusammenhangs zwischen der im Rahmen dieser Arbeit in der Wilden Weinrebe identifizierten Variante von BRE1-like 1 und einer möglichen Resistenz gegenüber *Botrytis cinerea* bedarf es jedoch weiterer Experimente. Die einzelnen Individuen der untersuchten Pools sollten im Weiteren auf ihr spezifisches

Verhalten gegenüber *Botrytis cinerea* getestet werden. Eine parallele Sangersequenzierung des Locus in den einzelnen Individuen ermöglicht bestenfalls eine Assoziation zwischen Phäno- und Genotyp und auf diese Weise die Identifizierung einer ursächlichen Mutation. Diese könnte dann wiederum durch *knock-in*-Versuche in nicht-resistenten Weinreben bestätigt werden sowie als molekularer Marker während der Einkreuzung der Resistenzeigenschaft in die Kulturrebe dienen.

Abkürzungsverzeichnis

°C	Grad Celsius
%	Prozent
A	Adenin
abs.	absolut
AP	Alkalische Phosphatase
AS	Aminosäure
ASCII	*American Standard Code for Information Interchange*
BLAST	*Basic Local Alignment Search Tool*
bp	Basenpaare
bzw.	beziehungsweise
C	Cytosin
ca.	circa
CDS	*coding DNA sequence*
CI	Chloroform-Isoamylalkohol
cm	Zentimeter
cpDNA	Chloroplasten-DNA
CTAB	Cetyltrimethylammoniumbromid
CWR	*crop wild relative*
ddPCR	*droplet digital PCR*
DMF	Dimethylformamid
d_N	Rate der nicht-synonymen Substitutionen
DNA	Desoxyribonukleinsäure
dNTP	Desoxyribonukleosidtriphosphat
d_S	Rate der synonymen Substitutionen
ds	doppelsträngig
DUF	Domäne unbekannter Funktion
E. coli	*Escherichia coli*
EDTA	Ethylendiamintetraacetat
E-Puffer	Elektrophorese-Puffer
et al.	et alii / aliae / alia
EtBr	Ethidiumbromid
EtOH	Ethanol
F-Allel	weibliches Allel des Geschlechtslocus

FAO	*Food and Agriculture Organization of the United Nations*
for	*forward* (vorwärts)
Fr	Frankreich
F_{ST}	Fixierungsindex
G	gefilterte Daten
g	Gramm
g	Erdbeschleunigung
G	Guanin
Gb	Gigabasen
GO	*gene ontology*
GP	Genpool
Gret1	*Grapevine Retrotransposon* 1
H_P	gepoolte Heterozygosität
HPLC	*high performance liquid chromatography*
Hyb	Hybridrebe
ID	Identifikationsnummer
Indel	Insertions-Deletions-Polymorphismus
iPBS	*inter-PBS amplification*
IPTG	Isopropyl-β-D-thiogalactopyranosid
IRAP	*inter-retrotransposon amplification polymorphism*
kb	Kilobasen
Ke	Ketsch
l	Liter
L.	Linné
LB-Medium	*lysogeny broth* Medium
LD	*linkage disequilibrium* (Kopplungsungleichgewicht)
Leu	Leucin
LINE	*long interspersed nuclear element*
log	Logarithmus
LTR	*long terminal repeat*
µ	Mittelwert
M	Molar
mA	Milliampère
MAF	minore Allelfrequenz
Mb	Megabasen

MCS	*multiple cloning site* (multiple Klonierungsstelle)
mg	Milligramm
MH-Allel	männliches oder hermaphroditisches Allel des Geschlechtslocus
min	Minuten
Mio.	Millionen
miRNA	micro RNA
µl	Mikroliter
ml	Milliliter
mM	Millimolar
mtDNA	Mitochondrien-DNA
N	undefinierte Base
NCBI	*National Center for Biotechnology Information*
ncRNA	nichtkodierende RNA
ng	Nanogramm
NGS	*Next-Generation Sequencing*
nm	Nanometer
OIV	*Organisation Internationale de la Vigne et du Vin*
p	p-Wert (Signifikanzwert)
PCR	*polymerase chain reaction* (Polymerasekettenreaktion)
pers.	persönliche
PGR	pflanzliche genetische Ressource
pH	potentia hydrogenii
Phe	Phenylalanin
Pi	Pisa
pmol	Picomol
PN40024	Pinot Noir Klon 40024
Prof.	Professor
PVPP	Polyvinylpolypyrrolidon
qPCR	quantitative PCR
QTL	*quantitative trait loci*
R	Rohdaten
r	Pearson-Korrelationskoeffizient
RAD	*restriction-site associated DNA*
rev	*reverse* (rückwärts)
RNA	Ribonukleinsäure

RPKM	*reads per kilobase per million reads*
RT	Raumtemperatur
σ	Standardabweichung
Ser	Serin
SINE	*short interspersed nuclear element*
SNP	*single nucleotide polymorphism* (Einzelnukleotidpolymorphismus)
Sp	Spanien
SRA	*sequence read archive*
SSR	*simple sequence repeats*
subsp.	Subspezies (Unterart)
T	Thymin
TBE	Tris-Borat-EDTA
Thr	Threonin
Ti/Tv	Verhältnis von Transitions- zu Transversionsmutationen
tRNA	Transfer-RNA
U	*unit* (Enzymeinheit)
u. a.	unter anderem
UTR	untranslatierte Region
UV	ultraviolett
v. Chr.	vor Christus
v/v	volume per volume (Volumenprozent)
VE	vollentsalzt
vgl.	vergleiche
Vol.	Volumen
VS	*Vitis vinifera* subsp. *sylvestris*
VV	*Vitis vinifera* subsp. *vinifera*
WGD	*whole genome duplication* (genomweite Duplikation)
WGS	*whole genome shotgun*-Strategien
X-Gal	5-Brom-4-chlor-3-indoxyl-β-D-galactopyranosid
z. B.	zum Beispiel
ZH_P	Z-transformierte gepoolte Heterozygosität

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang

Die hier angegebenen und beschriebenen Dateien befinden sich im beigefügten elektronischen Anhang.

- Digitale Version der hier vorliegenden Arbeit im PDF-Format
- Bilddateien aller Abbildungen dieser Arbeit im JPG-Format
- Alle verwendeten Perl-Skripte
- Referenzsequenzen der zweiten Runde der Poolseq-Kartierung im FASTA-Format
- Sequenzen der anhand ihrer signifikant reduzierten Diversität identifizierten Signale im FASTA-Format
- S1 - S4: *Variant tables* der vier Pools im TXT-Format (Listen der identifizierten Polymorphismen inkl. der Anzahl und Sequenz der zugrundeliegenden Allele, deren Abdeckung in der Kartierung sowie Allelfrequenzen)
- S5: Experimentelle Validierung zufällig ausgewählter SNPs
- S6 - S9: Mit dem *sliding window* erhobene H_P- und ZH_P-Werte der vier Pools
- S10: Identifizierte Kandidatenregionen der vier Pools mit detaillierten Positions- und Längeninformationen
- S11: Tabelle der identifizierten Kandidatengene
- S12: Tabelle der Kandidatengene, denen ein Züchtungs-relevanter *GO-Term* zugeordnet wurde

Literatur

Adams, K.L. & Wendel, J.F., 2005. Polyploidy and genome evolution in plants. *Current opinion in plant biology*, 8(2), pp.135–41.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–10.

Altshuler, D. et al., 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407(6803), pp.513–6.

Amaral, A.J. et al., 2011. Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PloS one*, 6(4), p.e14782.

Anderson, P.K. et al., 2004. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends in Ecology & Evolution*, 19(10), pp.535–544.

De Andrés, M.T. et al., 2012. Genetic diversity of wild grapevine populations in Spain and their genetic relationships with cultivated grapevines. *Molecular ecology*, 21(4), pp.800–16.

Antcliff, A.J., 1980. Inheritance of sex in Vitis. *Annales de l'amélioration des plantes*, 30(2), pp.113–122.

Anzani, R. et al., 1990. Wild grapevine (Vitis vinifera var . silvestri) in ltaly : Distribution , characteristics and germplasm preservatlon - 1989 report. *Vitis*, 29, pp.97–113.

Aradhya, M. et al., 2008. Genetic Structure, Differentiation, and Phylogeny of the Genus *Vitis*: Implications for Genetic Conservation. *Proceedings of the Fifth International Symposium on the Taxonomy of Cultivated Plants*, 799(2008), pp.43–49.

Aradhya, M.K. et al., 2003. Genetic structure and differentiation in cultivated grape, Vitis vinifera L. *Genetical research*, 81(3), pp.179–92.

Arnold, C., 2002. Ecologie de la vigne sauvage (Vitis vinifera L. ssp. silvestris (Gmelin) Hegi.) dans les forêts alluviales et colluviales d'Europe. *Geobotanica Helvetica*, 76, p.256.

Arnold, C. et al., 2005. Is there a future for wild grapevine (Vitis vinifera subsp. silvestris) in the Rhine Valley? *Biodiversity and Conservation*, 14(6), pp.1507–1523.

Arnold, C., Gillet, F. & Gobat, J.M., 1998. Situation de la vigne sauvage Vitis vinifera ssp. silvestris en Europe. *Vitis*, 37(4), pp.159–170.

Arroyo-García, R. et al., 2006. Multiple origins of cultivated grapevine (Vitis vinifera L. ssp.

sativa) based on chloroplast DNA polymorphisms. *Molecular ecology*, 15(12), pp.3707–14.

Arroyo-García, R.A. & Revilla, E., 2013. The Current Status of Wild Grapevine Populations (Vitis vinifera ssp sylvestris) in the Mediterranean Basin. In B. Sladonja, ed. *The Mediterranean Genetic Code - Grapevine and Olive*. pp. 51–72.

Axelsson, E. et al., 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*, (2418), pp.1–6.

Bacilieri, R. & This, P., 2007. GrapeGen06, an European Project for the Management and Conservation of Grapevine Genetic Resources. Available at: https://www1.montpellier.inra.fr/grapegen06/ [Accessed March 30, 2017].

El Baidouri, M. & Panaud, O., 2013. Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome biology and evolution*, 5(5), pp.954–65.

Bao, W., Kojima, K.K. & Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1), p.11.

Barnaud, A. et al., 2010. Linkage disequilibrium in wild French grapevine, Vitis vinifera L. subsp. silvestris. *Heredity*, 104(5), pp.431–7.

Barth, S. et al., 2009. Genotypes and phenotypes of an ex situ Vitis vinifera ssp. sylvestris (Gmel.) beger germplasm collection from the upper rhine valley. *Genetic Resources and Crop Evolution*, 56(8), pp.1171–1181.

Benjak, A., Forneck, A. & Casacuberta, J.M., 2008. Genome-wide analysis of the “cut-and-paste” transposons of grapevine. *PloS one*, 3(9), p.e3107.

Bennetzen, J.L., 2000. Transposable element contributions to plant genome evolution. *Plant Mol Biol*, 42, pp.251–269.

Berkelmann-Löhnertz, B. et al., 2008. Ohne Kupfer geht es nicht – Status quo im ökologischen Weinbau nach vier Jahren BÖL-Verbundprojekt Fachgespräch „Bedeutung von Kupfer für den Pflanzenschutz, insbesondere für den Ökologischen Landbau – Reduktions- und Ersatzstrategien“, Berlin 29.01.2008. , pp.17–20.

Bevan, M. et al., 1997. Objective: The complete sequence of a plant genome. *Plant Cell*, 9(4), pp.746–748.

Biagini, B. et al., 2014. Italian wild grapevine (Vitis vinifera L. subsp. sylvestris) population: insights into eco-geographical aspects and genetic structure. *Tree Genetics & Genomes*, pp.1369–1385.

Bilz, M. et al., 2011. *European Red List of Vascular Plants*, Available at: http://data.iucn.org/dbtw-wpd/edocs/RL-4-016.pdf.

Boavida, L.C. et al., 2009. A collection of Ds insertional mutants associated with defects in male gametophyte development and function in Arabidopsis thaliana. *Genetics*, 181(4), pp.1369–85.

Bodor, P. et al., 2010. Conservation value of the native Hungarian wild grape (Vitis sylvestris Gmel.) evaluated by microsatellite markers. *Vitis - Journal of Grapevine Research*, 49(1), pp.23–27.

Boffey, S.A. & Leech, R.M., 1982. Chloroplast DNA Levels and the Control of Chloroplast Division in Light-Grown Wheat Leaves. *Plant Physiology*, 69(6), pp.1387–1391.

Bolger, M.E. et al., 2014. Plant genome sequencing - applications for crop improvement. *Current Opinion in Biotechnology*, 26, pp.31–37.

Bowers, J. et al., 1999. Historical Genetics: The Parentage of Chardonnay, Gamay, and Other Wine Grapes of Northeastern France. *Science (New York, N.Y.)*, 285(5433), pp.1562–1565.

Brasnyó, P. et al., 2011. Resveratrol improves insulin sensitivity, reduces oxidative stress and activates the Akt pathway in type 2 diabetic patients. *British Journal of Nutrition*, 106(3), pp.383–389.

Braverman, J.M. et al., 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, 140(2), pp.783–796.

Brown, A.G., 1975. Apples. In J. Janick & I. N. Moore, eds. *Advances in fruit breeding*. West Lafayette: Purdue Univ. Press, pp. 3–37.

Brozynska, M., Furtado, A. & Henry, R.J., 2015. Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant biotechnology journal*, pp.1–16.

Buehler, B. et al., 2010. Rapid quantification of DNA libraries for next-generation sequencing. *Methods*, 50(4), pp.S15–S18.

Canaguier, A. et al., 2014. Improvement of the Grape Genome Assembly. In *Poster anlässlich der Tagung "Plant and Animal Genome XXII."* San Diego, CA, USA.

Caporali, E. et al., 2003. The arrest of development of useless reproductive organs in the unisexual flower of Vitis vinifera ssp silvestris. *Acta Horticulturae*, 603, pp.225–228.

Carlson, C.S. et al., 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, 15(11), pp.1553–1565.

Carneiro, M. et al., 2014. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science (New York, N.Y.)*, 345(6200), pp.1074–9.

Casacuberta, E. & González, J., 2013. The impact of transposable elements in environmental adaptation. *Molecular Ecology*, 22(6), pp.1503–1517.

Castro, I. et al., 2012. Effectiveness of AFLPs and retrotransposon-based markers for the identification of Portuguese grapevine cultivars and clones. *Molecular biotechnology*, 52(1), pp.26–39.

Charlesworth, B., Morgan, M.T. & Charlesworth, D., 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134, pp.1289–1303.

Charlesworth, D., 2006. Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLoS Genetics*, 2(4), pp.379–384.

Choi, J.W. et al., 2015. Whole-genome resequencing analyses of five pig breeds, including Korean wild and native, and three European origin breeds. *DNA Research*, 22(4), pp.259–267.

Claros, M.G. et al., 2012. Why Assembling Plant Genome Sequences Is So Challenging. *Biology*, 1(2), pp.439–459.

Cook, E.H. & Scherer, S.W., 2008. Copy-number variations associated with neuropsychiatric conditions. *Nature*, 455(October), pp.919–923.

Cummins, I.N. & Aldwinckle, H.S., 1979. Breeding Tree Crops. In H. M.K., ed. *Biology and breeding for resistance to arthropods and pathogens in agricultural plants*. Texas Agricultural Experiment Station: Texas A and M University, pp. 528–545.

CWRIS, 2005. PGR Forum Crop Wild Relative Information System. Available at: http://www.pgrforum.org/cwris/cwris.asp [Accessed October 19, 2015].

D'Onofrio, C. et al., 2010. Retrotransposon-based molecular markers for grapevine species and cultivars identification. *Tree Genetics and Genomes*, 6(3), pp.451–466.

Dalbó, M.A. et al., 2000. A gene controlling sex in grapevines placed on a molecular marker-based genetic map. *Genome*, 43(2), pp.333–40.

Darwin, C.R. & Wallace, A.R., 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *Journal of the proceedings of the Linnean Society of London. Zoology*, 3(9), pp.45–62.

Davies, P.A. & Gray, G., 2002. Long-range PCR. *Methods in molecular biology*, 187, pp.51–55.

Denoeud, F. et al., 2014. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science (New York, N.Y.)*, 345(6201), pp.1181–4.

Dhawan, R. et al., 2009. HISTONE MONOUBIQUITINATION1 Interacts with a Subunit of the Mediator Complex and Regulates Defense against Necrotrophic Fungal Pathogens in Arabidopsis. *THE PLANT CELL ONLINE*, 21(3), pp.1000–1019.

Diesner, M.-O. et al., 2014. Kupfer im Bio-Landbau: Hintergrund, Herausforderungen und

Handlungsempfehlungen. *Öko-Institut e.V.*, pp.1–35.

Dietzen, S., 2012. *Genetische Variabilität im Vergleich der Kulturrebe zur Wilden Weinrebele*. Johannes-Gutenberg Universität Mainz.

Ding, B., 2016. *Vergleichende Analyse der Blütentranskriptome von Vitis vinifera ssp. sylvestris und Vitis vinifera ssp. vinifera*. Johannes Gutenberg-Universität Mainz.

Dobritsa, A. a. et al., 2011. A large-scale genetic screen in Arabidopsis to identify genes involved in pollen exine production. *Plant physiology*, 157(2), pp.947–70.

Dohm, J.C. et al., 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research*, 36(16), p.e105.

Don, R.H. et al., 1991. "Touchdown" PCR to circumvent spurious priming during gene amplification. *Nucleic acids research*, 19(14), p.4008.

Donald, T.M. et al., 2002. Identification of resistance gene analogs linked to a powdery mildew resistance locus in grapevine. *TAG Theoretical and Applied Genetics*, 104(4), pp.610–618.

Duan, D. et al., 2016. An ancestral allele of grapevine transcription factor MYB14 promotes plant defence. *Journal of Experimental Botany*.

Duarte, J.M. et al., 2010. Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology*, 10(1), p.61.

Dwivedi, S.L. et al., 2008. Enhancing Crop Gene Pools with Beneficial Traits Using Wild Relatives. *Plant Breeding Reviews*, 30, pp.179–230.

Ekhvaia, J. & Akhalkatsi, M., 2010. Morphological variation and relationships of Georgian populations of Vitis vinifera L. subsp. sylvestris (C.C. Gmel.) Hegi. *Flora: Morphology, Distribution, Functional Ecology of Plants*, 205(9), pp.608–617.

Elad, Y. et al., 2007. Botrytis ssp. and Diseases They Cause in Agricultural Systems - An Introduction. In Y. Elad et al., eds. *Botrytis: Biology, Pathology and Control*. Dordrecht: Springer, pp. 1–8.

Ellegren, H., 2004. Microsatellites: simple sequences with complex evolution. *Nature Reviews*, 5(June), pp.435–445.

Emberton, J. et al., 2005. Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries. *Genome Research*, 15(10), pp.1441–1446.

English, A.C. et al., 2012. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS ONE*, 7(11), p.e47768.

Esquinas-Alcázar, J.T., 1993. Plant genetic resources. In M. D. Hayward et al., eds. *Plant Breeding*. Dordrecht: Springer Netherlands, pp. 33–51.

Fagny, M. et al., 2014. Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. *Molecular Biology and Evolution*, 31(7), pp.1850–1868.

FAO, 2013. *Food and Agricultural commodities production 2013*, Available at: http://faostat3.fao.org/browse/rankings/commodities_by_regions/E.

Fechter, I. et al., 2012. Candidate genes within a 143 kb region of the flower sex locus in Vitis. *Molecular genetics and genomics (MGG)*, 287(3), pp.247–59.

Finnegan, D.J., 1989. Eukaryotic transposable elements and genome evolution. *Trends in genetics : TIG*, 5(4), pp.103–7.

Fischer, S., 2012. *Vergleichende Analyse der Genome der Wildrebe und der Kulturrebe mittels Hochdurchsatzsequenzierung*. Johannes Gutenberg-Universität Mainz.

Fleischmann, A. et al., 2014. Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Annals of Botany*, 114(8), pp.1651–1663.

Ford-Lloyd, B. V. et al., 2011. Crop Wild Relatives—Undervalued, Underutilized and under Threat? *BioScience*, 61(7), pp.559–565.

Francia, E. et al., 2005. Marker assisted selection in crop plants. *Plant Cell, Tissue and Organ Culture*, 82(3), pp.317–342.

Di Genova, A. et al., 2014. Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC plant biology*, 14, p.7.

German, J.B. & Walzem, R.L., 2000. The health benefits of wine. *Annual review of nutrition*, 20, pp.561–593.

Gheyas, A.A. et al., 2014. Functional classification of 15 million SNPs detected from diverse chicken populations. *DNA Research*, 22(3), pp.205–217.

Goremykin, V. V. et al., 2008. Mitochondrial DNA of Vitis vinifera and the Issue of Rampant Horizontal Gene Transfer. *Molecular Biology and Evolution*, 26(1), pp.99–110.

Govaerts, R., 2001. How Many Species of Seed Plants Are There? *Taxon*, 50(4), pp.1085–1090.

Grant, S.G. et al., 1990. Differential plasmid rescue from transgenic mouse DNAs into Escherichia coli methylation-restriction mutants. *Proceedings of the National Academy of Sciences of the United States of America*, 87(June), pp.4645–4649.

Grassi, F. et al., 2003. Evidence of a secondary grapevine domestication centre detected by SSR analysis. *Theoretical and Applied Genetics*, 107(7), pp.1315–1320.

Grassi, F. et al., 2008. Historical isolation and Quaternary range expansion of divergent lineages in wild grapevine. *Biological Journal of the Linnean Society*, 95(3), pp.611–619.

Grassi, F. et al., 2006. Phylogeographical structure and conservation genetics of wild grapevine. *Conservation Genetics*, 7(6), pp.837–845.

Green, P., 1997. Against a whole-genome shotgun. *Genome Res.*, 7(206), pp.410–417.

Griffith, A.J. et al., 2000. Spontaneous mutations. In *An Introduction to Genetic Analysis*. New York: W. H. Freeman.

Grossman, S.R. et al., 2010. A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. *Science*, 327(5967), pp.883–886.

Guo, X. et al., 2016. Whole-genome resequencing of Xishuangbanna fighting chicken to identify signatures of selection. *Genetics Selection Evolution*, 48(1), p.62.

Gupta, P.K., Roy, J.K. & Prasad, M., 2001. Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science*, 80(4), pp.524–535.

Hajjar, R. & Hodgkin, T., 2007. The use of wild relatives in crop improvement: a survey of developments over the last 20 years. *Euphytica*, 156(1–2), pp.1–13.

Hamilton, J.P. & Robin Buell, C., 2012. Advances in plant genome sequencing. *The Plant Journal*, 70(1), pp.177–190.

Hammer, K., 1984. Das Domestikationssyndrom. *Die Kulturpflanze*, 32(1), pp.11–34.

Harlan, J.R. & De Wet, J.M.J., 1971. Toward a Rational Classification of Cultivated Plants. *Taxon*, 20(4), pp.509–517.

Hayward, A.C. et al., 2015. Molecular Marker Applications in Plants. In J. Batley, ed. *Plant Genotyping: Methods and Protocols*. New York: Springer Science+Business Media, pp. 13–27.

He et al., 1981. The inheritance of some characteristics in interspecific hybrid of grape, with special reference to cold resistance. *Acta Hortic.*, 8(1), p.18.

Hegarty, M.J. & Hiscock, S.J., 2008. Genomic Clues to the Evolutionary Success of Polyploid Plants. *Current Biology*, 18(10), pp.R435–R444.

Hegi, G., 1925. *Illustrierte Flora von Mitteleuropa. Band 5.*, München: J. F. Lehmanns Verlag.

Henry, R.J., 2012. Next-generation sequencing for understanding and accelerating crop domestication. *Briefings in functional genomics*, 11(1), pp.51–6.

Hewitt, G., 1999. Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, 68(1–2), pp.87–112.

Heywood, V. et al., 2007. Conservation and sustainable use of crop wild relatives. *Agriculture, Ecosystems & Environment*, 121(3), pp.245–255.

Hilu, K.W., 1993. Polyploidy and the origin of Domesticated Plants. *Evolution*, 80(12), pp.1494–1499.

Hirsch, S., Baumberger, R. & Grossniklaus, U., 2012. Epigenetic Variation, Inheritance, and Selection in Plant Populations. *Cold Spring Harbor Symposia on Quantitative Biology*, 77, pp.97–104.

Huang, J. et al., 2016. Control of Anther Cell Differentiation by the Small Protein Ligand TPD1 and Its Receptor EMS1 in Arabidopsis. *PLoS genetics*, 12(8), p.e1006147.

Huang, S. et al., 2009. The genome of the cucumber, Cucumis sativus L. *Nature genetics*, 41(12), pp.1275–1281.

Hufford, M.B. et al., 2012. Comparative population genomics of maize domestication and improvement. *Nature genetics*, 44(7), pp.808–11.

Hurst, G.D.D. & Werren, J.H., 2001. The role of selfish genetic elements in eukaryotic evolution. *Nature Reviews Genetics*, 2(8), pp.597–606.

Hwang, W.W. et al., 2003. A conserved RING finger protein required for histone H2B monoubiquitination and cell size control. *Molecular cell*, 11(1), pp.261–6.

Ibarra-Laclette, E. et al., 2013. Architecture and evolution of a minute plant genome. *Nature*, 498(7452), pp.94–8.

Imelfort, M. & Edwards, D., 2009. De novo sequencing of plant genomes using second-generation technologies. *Briefings in Bioinformatics*, 10(6), pp.609–618.

ITPGRFA, 2009. International Treaty on Plant Genetic Resources for Food and Agriculture.

Jablonka, E. & Raz, G., 2009. Transgenerational Epigenetic Inheritance: Prevalence, Mechanisms, and Implications for the Study of Heredity and Evolution. *The Quarterly Review of Biology*, 84(2), pp.131–176.

Jaillon, O. et al., 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161), pp.463–7.

Jiang, J. & Gill, B.S., 2006. Current status and the future of fluorescence in situ hybridization (FISH) in plant genome research. *Genome*, 49(9), pp.1057–1068.

Jin, W. et al., 2009. Improved cold-resistant performance in transgenic grape (Vitis vinifera L.) overexpressing cold-inducible transcription factors AtDREB1b. *HortScience*, 44(1), pp.35–39.

Johnsson, M. et al., 2016. Feralisation targets different genomic loci to domestication in the chicken. *Nature communications*, 7, p.12950.

Kalendar, R. et al., 2008. Cassandra retrotransposons carry independently transcribed 5S RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 105(15), pp.5833–8.

Kamberov, Y.G. et al., 2013. Modeling Recent Human Evolution in Mice by Expression of a Selected EDAR Variant. *Cell*, 152(4), pp.691–702.

Kaneko, Y. & Bang, S.W., 2014. Interspecific and intergeneric hybridization and chromosomal engineering of Brassicaceae crops. *Breeding Science*, 64(1), pp.14–22.

Kaplan, N.L., Hudson, R.R. & Langley, C.H., 1989. The “hitchhiking effect” revisited. *Genetics*, 123(4), pp.887–99.

Kellogg, E.A. & Bennetzen, J.L., 2004. The Evolution of Nuclear Genome Structure in Seed Plants. *American Journal of Botany*, 91(10), pp.1709–1725.

Kibbe, W.A., 2007. OligoCalc: An online oligonucleotide properties calculator. *Nucleic Acids Research*, 35, pp.43–46.

Kim, Y. & Stephan, W., 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2), pp.765–777.

Kimura, M., 1983. *The Neutral Theory of Molecular Evolution*, Cambridge: Cambridge University Press.

King, I. et al., 1997. Introgression of salt-tolerance genes from Thinopyrum hessarahicum into wheat. *New Phytologist*, 137, pp.75–81.

Kobayashi, S., Goto-Yamamoto, N. & Hirochika, H., 2004. Retrotransposon-induced mutations in grape skin color. *Science (New York, N.Y.)*, 304(5673), p.982.

Koes, R., Verweij, W. & Quattrocchio, F., 2005. Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. *Trends in plant science*, 10(5), pp.236–42.

Kubo, T. & Newton, K.J., 2008. Angiosperm mitochondrial genomes and mutations. *Mitochondrion*, 8(1), pp.5–14.

Kühne, S. et al., 2009. Anwendung kupferhaltiger Pflanzenschutzmittel in Deutschland. *Journal für Kulturpflanzen*, 61(4), pp.126–130.

Lamichhaney, S. et al., 2016. A beak size locus in Darwin's finches facilitated character displacement during a drought. *Science (New York, N.Y.)*, 352(6284), pp.470–4.

Lamppa, G.K., Elliot, L. V. & Bendich, A.J., 1980. Changes in chloroplast number during pea leaf development. *Planta*, 148(5), pp.437–443.

Laurie, M.T. et al., 2013. Simultaneous digital quantification and fluorescence-based size characterization of massively parallel sequencing libraries. *BioTechniques*, 55(2).

Ledesma-Krist, G., Schumann, F. & Maul, E., 2015. Die Wildrebenpopulation auf der Rheininsel Ketsch - Eine wertvolle genetische Ressource. In *Deutsches Weinbaujahrbuch 2015*. Stuttgart, pp. 106–118.

Ledesma-Krist, G.M. et al., 2013. Überlebenssicherung der Wildrebe Vitis vinifera L. ssp. sylvestris (C. C. Gmel.) Hegi in den Rheinauen durch gezieltes in situ-Management. *Abschlussbericht 2008-2013*, pp.1–94. Available at: https://service.ble.de/fpd_ble/index2.php?detail_id=690&site_key=151&stichw_suche=Wildrebe&zeilenzahl_zaehler=6.

Levadoux, L., 1956. Les populations sauvages et cultivées de Vitis vinifera L. *Annales de l'amélioration des plantes*, (6), pp.59–118.

Li, R. et al., 2010. The sequence and de novo assembly of the giant panda genome. *Nature*, 463(7279), pp.311–7.

Lijavetzky, D. et al., 2007. High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. *BMC genomics*, 8, p.424.

Lijavetzky, D. et al., 2006. Molecular genetics of berry colour variation in table grape. *Molecular genetics and genomics : MGG*, 276(5), pp.427–35.

Ling, H.-Q. et al., 2013. Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature*, 496(7443), pp.87–90.

Lisch, D., 2013. How important are transposons for plant evolution? *Nature reviews. Genetics*, 14(1), pp.49–61.

Liu, H.T. et al., 2008. The calmodulin-binding protein kinase 3 is part of heat-shock signal transduction in Arabidopsis thaliana. *Plant Journal*, 55(5), pp.760–773.

Lodhi, M.A. & Reisch, B.I., 1995. Nuclear DNA content of Vitis species , cultivars , and other genera of the Vitaceae. *Theor Appl Genet*, 90, pp.11–16.

Logan, D.C., 2007. *Plant Mitochondria*, Jon Wiley & Sons: Blackwell Publishing's Annual Plant Reviews.

Lopes, M.S. et al., 2009. New insights on the genetic basis of Portuguese grapevine and on

grapevine domestication. *Genome*, 52(9), pp.790–800.

Magi, A. et al., 2012. Read count approach for DNA copy number variants detection. *Bioinformatics*, 28(4), pp.470–478.

Marcussen, T. et al., 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science (New York, N.Y.)*, 345(6194), p.1250092.

Marrano, A. et al., 2017. SNP-Discovery by RAD-Sequencing in a Germplasm Collection of Wild and Cultivated Grapevines (V. vinifera L.) S. Amancio, ed. *PLOS ONE*, 12(1), p.e0170655.

Mascagni, F. et al., 2015. Repetitive DNA and Plant Domestication: Variation in Copy Number and Proximity to Genes of LTR-Retrotransposons among Wild and Cultivated Sunflower (Helianthus annuus) Genotypes. *Genome Biology and Evolution*, 7(12), pp.3368–3382.

De Mattia, F. et al., 2008. Study of genetic relationships between wild and domesticated grapevine distributed from Middle East Regions to European countries. *Rendiconti Lincei*, 19(3), pp.223–240.

Matus, J.T., Aquea, F. & Arce-Johnson, P., 2008. Analysis of the grape MYB R2R3 subfamily reveals expanded wine quality-related clades and conserved gene structure organization across Vitis and Arabidopsis genomes. *BMC plant biology*, 8, p.83.

Maul, E., 2016. Die genetische Ressourcen der Rebe: Erhaltungssituation in Europa. In *Jahrbuch 2015 der Braunschweigischen Wissenschaftlichen Gesellschaft*. Braunschweig: J. Cramer Verlag, pp. 345–358.

Maxted, N. et al., 2006. Towards a definition of a crop wild relative. *Biodiversity and Conservation*, 15(8), pp.2673–2685.

Maxted, N. & Kell, S., 2012. PGR Secure: enhanced use of traits from crop wild relatives and landraces to help adapt crops to climate change. *Crop Wild Relative*, 8(8), pp.4–7.

Maxted, N. & Kell, S.P., 2009. Establishment of a Global Network for the in situ conservation of crop wild relatives: status and needs. *Situ Conservation of Crop Wild Relatives: Status and Needs. FAO Commission on Genetic Resources for Food and Agriculture, Rome, Italy*, p.266. Available at: http://www.pgrforum.org/documents/Global_in_situ_CWR_conservation_network.pdf.

Maynard Smith, J. & Haigh, J., 1974. The hitch-hiking effect of a favourable gene. *Genetical research*, 23(1), pp.23–35.

McClintock, B., 1984. The significance of responses of the genome to challenge. *Science (New York, N.Y.)*, 226(4676), pp.792–801.

McGovern, P.E. et al., 1996. Neolithic resinated wine. *Nature*, 381(6582), pp.480–481.

Meacham, F. et al., 2011. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, 12(1), p.451.

Meilleur, B. & Hodgkin, T., 2004. In situ conservation of crop wild relatives: statuts and trends. *Biodiversity and Conservation*, 13, pp.663–684.

Meyers, B.C., 2003. Genome-Wide Analysis of NBS-LRR-Encoding Genes in Arabidopsis. *THE PLANT CELL ONLINE*, 15(4), pp.809–834.

Michael, T.P., 2014. Plant genome size variation: bloating and purging DNA. *Briefings in Functional Genomics*, 13(4), pp.308–317.

Michael, T.P. & VanBuren, R., 2015. Progress, challenges and the future of crop genomes. *Current Opinion in Plant Biology*, 24, pp.71–81.

Miller, J.R., Koren, S. & Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6), pp.315–27.

Montague, M.J. et al., 2014. Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. *Proceedings of the National Academy of Sciences of the United States of America*, 111(48), pp.17230–5.

Mortazavi, A. et al., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7), pp.621–8.

Moses, A.M., 2009. Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. *BMC Evolutionary Biology*, 9(1), p.286.

Mozo, T. et al., 1998. Construction and characterization of the IGF Arabidopsis BAC library. *Molecular and General Genetics MGG*, 258(5), pp.562–570.

Mullins, M.G., Bouquet, A. & Williams, L.E., 1992. *Biology of the grapevine*, Cambridge: Cambridge University Press.

Mullis, K. et al., 1986. Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, 51, pp.263–273.

Myles, S. et al., 2011. Genetic structure and domestication history of the grape. *Proceedings of the National Academy of Sciences of the United States of America*, 108(9), pp.3530–5.

Naito, K. et al., 2006. Dramatic amplification of a rice transposable element during recent domestication. *Proceedings of the National Academy of Sciences of the United States of America*, 103(47), pp.17620–5.

Nault, L.R. et al., 1982. Response of annual and perennial teosintes zea to 6 maize viruses.

Plant Disease. 1982;, 66(1), p.61–62.

Neale, D.B. et al., 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, 15(3), p.R59.

Negrul, A.M., 1938. Evolution of cultivated forms of grapes. *CR Acad Sci USSR*, 18, pp.585–588.

Nei, M., 2005. Selectionism and neutralism in molecular evolution. *Molecular Biology and Evolution*, 22(12), pp.2318–2342.

Nguyen, B.D. et al., 2003. Identification and mapping of the QTL for aluminum tolerance introgressed from the new source, Oryza Rufipogon Griff., into indica rice (Oryza sativa L.). *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 106(4), pp.583–593.

Nick, P., 2012. Von der Ex-situ-Erhaltung bis zur Nutzung in der nachhaltigen Landwirtschaft: Das Beispiel der Europäischen Wildrebe. *Berichte Ges Pflanzenbauwiss*, 6, pp.36–38.

Nickrent, D.L. & Soltis, D.E., 1995. A Comparison of Angiosperm Phylogenies from Nuclear 18S rDNA and rbcL Sequences. *Annals of the Missouri Botanical Garden*, 82(2), p.208.

Nicolas, S.D. et al., 2016. Genetic diversity, linkage disequilibrium and power of a large grapevine (Vitis vinifera L) diversity panel newly designed for association studies. *BMC Plant Biology*, 16(1), p.74.

Nielsen, R., 2005. Molecular signatures of natural selection. *Annual review of genetics*, 39, pp.197–218.

Ocete, R. et al., 2008. Comparative analysis of wild and cultivated grapevine (Vitis vinifera) in the Basque Region of Spain and France. *Agriculture, Ecosystems and Environment*, 123(1–3), pp.95–98.

Ocete, R. & Lara, R., 1994. Consideraciones sobre la ausencia de síntomas de ataque por filoxera en poblaciones autóctonas de Vitis vinifera silvestris (Gmelin) Hegi. *Bol. San. Veg. Plagas*, (20), pp.631–636.

Ogbonnaya, F.C. et al., 2013. Synthetic hexaploids: Harnessing species of the primary gene pool for wheat improvement. *Plant Breeding Reviews*, 37(1), pp.35–122.

OIV, 2015. *World vitiviniculture situtation*, Available at: http://www.oiv.int/oiv/files/Report_Mainz_Congress_2015_OIV_EN.pdf.

Oleksyk, T.K., Smith, M.W. & O'Brien, S.J., 2010. Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537), pp.185–205.

Olmo, H.P., 1976. Grapes. In N. W. Simmonds, ed. *Evolution of Crop Plants*. London: Longman, pp. 294–298.

Osborn, T.C. et al., 2007. Insights and Innovations from Wide Crosses: Examples from Canola and Tomato. *Crop Science*, 47, pp.228–237.

Otto, D. et al., 2014. The columnar mutation ("Co gene") of apple (Malus × domestica) is associated with an integration of a Gypsy-like retrotransposon. *Molecular Breeding*, 33(4), pp.863–880.

Paux, E. et al., 2008. A Physical Map of the 1-Gigabase Bread Wheat Chromosome 3B. *Science*, 322(5898), pp.101–104.

Pavangadkar, K., Thomashow, M.F. & Triezenberg, S.J., 2010. Histone dynamics and roles of histone acetyltransferases during cold-induced gene regulation in Arabidopsis. *Plant molecular biology*, 74(1–2), pp.183–200.

Pellicer, J., Fay, M.F. & Leitch, I.J., 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, 164(1), pp.10–15.

Pelsy, F., 2010. Molecular and cellular mechanisms of diversity within grapevine varieties. *Heredity*, 104(4), pp.331–40.

Perry, G.H. et al., 2007. Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, 39(10), pp.1256–1260.

Picq, S. et al., 2014. A small XY chromosomal region explains sex determination in wild dioecious. *BMC plant biology*, 14(1), p.229.

PlaBi, 2015. Timeline of published plant genomes (2000-present). Available at: http://plabipd.de/timeline_view.ep [Accessed November 24, 2015].

Poptsova, M.S. et al., 2014. Non-random DNA fragmentation in next-generation sequencing. *Scientific Reports*, 4.

Price, H.J. et al., 2005. A Sorghum bicolor × S. macrospermum hybrid recovered by embryo rescue and culture. *Australian Journal of Botany*, 53(6), pp.579–58.

Proost, S. et al., 2011. Journey through the past: 150 million years of plant genome evolution. *The Plant journal : for cell and molecular biology*, 66(1), pp.58–65.

Proost, S. et al., 2015. PLAZA 3.0: An access point for plant comparative genomics. *Nucleic Acids Research*, 43(D1), pp.D974–D981.

Qanbari, S. et al., 2012. A high resolution genome-wide scan for significant selective sweeps: an application to pooled sequence data in laying chickens. *PloS one*, 7(11), p.e49525.

Quail, M.A. et al., 2008. A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, 5(12), pp.1005–1010.

Ramos, M.J. et al., 2014. Flower development and sex specification in wild grapevine. *BMC Genomics*, 15(1), p.1095.

Ranc, N. et al., 2008. A clarified position for solanum lycopersicum var. cerasiforme in the evolutionary history of tomatoes (solanaceae). *BMC Plant Biology*, 8(1), p.130.

Rellstab, C. et al., 2013. Validation of SNP Allele Frequencies Determined by Pooled Next-Generation Sequencing in Natural Populations of a Non-Model Plant Species G. A. Marais, ed. *PLoS ONE*, 8(11), p.e80422.

Riahi, L. et al., 2013. Characterization of single nucleotide polymorphism in Tunisian grapevine genome and their potential for population genetics and evolutionary studies. *Genetic Resources and Crop Evolution*, 60(3), pp.1139–1151.

Richards, E.J., 2011. Natural epigenetic variation in plant species: a view from the field. *Current Opinion in Plant Biology*, 14(2), pp.204–209.

Rick, C.M. & Chetelat, R.T., 1995. Utilization of related wild species for tomato improvement. *Acta Horticulturae*, (412), pp.21–38.

Rivera Núñez, D. et al., 2007. Multivariate analysis of Vitis subgenus Vitis seed morphology. *Vitis - Journal of Grapevine Research*, 46(4), pp.158–167.

Rubin, C.-J. et al., 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, 464(7288), pp.587–91.

Rusk, N., 2011. Torrents of sequence. *Nature Methods*, 8(1), pp.44–44.

Sabeti, P.C. et al., 2006. Positive Natural Selection in the Human Lineage. *Science*, 312(June), pp.1614–1620.

Salmaso, M. et al., 2004. Genome diversity and gene haplotypes in the grapevine (Vitis vinifera L.), as revealed by single nucleotide polymorphisms. *Molecular breeding : new strategies in plant improvement*, 14, pp.385–395.

Sambrook, J. & Russell, D.W., 2006. Agarose Gel Electrophoresis. *Cold Spring Harbor Protocols*, 2006, p.pdb.prot4020-prot4020.

Sato, S. et al., 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400), pp.635–41.

Saxena, K.B. et al., 2005. A cytoplasmic-nuclear male-sterility system derived from a cross between Cajanus cajanifolius and Cajanus cajan. *Euphytica*, 145(3), pp.289–294.

Schlötterer, C. et al., 2014. Sequencing pools of individuals — mining genome-wide

polymorphism data without big funding. *Nature Reviews Genetics*, 15(11), pp.749–763.

Schnable, P.S. et al., 2009. The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, 326(5956), pp.1112–1115.

Schröder, S. et al., 2015. Crop wild relatives as genetic resources – the case of the European wild grape. *Canadian Journal of Plant Science*, 95(5), pp.905–912.

Schwander, F. et al., 2012. Rpv10: a new locus from the Asian Vitis gene pool for pyramiding downy mildew resistance loci in grapevine. *Theoretical and Applied Genetics*, 124(1), pp.163–176.

Scott, N.S. & Possingham, J. V., 1980. Chloroplast DNA in expanding spinach leaves. *Journal of Experimental Botany*, 31(4), pp.1081–1092.

Sebolt, A.M., Shoemaker, R.C. & Diers, B.W., 2000. Analysis of a Quantitative Trait Locus Allele from Wild Soybean That Increases Seed Protein Concentration in Soybean. *Crop Science*, 40, pp.1438–1444.

Seiwert, M., 2014. *Fingerprinting bei der wilden Weinrebe*. Johannes Gutenberg-Universität Mainz.

Shearer, L.A. et al., 2014. Fluorescence in situ hybridization and optical mapping to correct scaffold arrangement in the tomato genome. *G3: Genes, Genomes, Genetics*, 4(8), pp.1395–405.

Shimazaki, M. et al., 2011. Pink-colored grape berry is the result of short insertion in intron of color regulatory gene. *PloS one*, 6(6), p.e21308.

Shulaev, V. et al., 2011. The genome of woodland strawberry (Fragaria vesca). *Nature Genetics*, 43(2), pp.109–116.

Da Silva, C. et al., 2013. The High Polyphenol Content of Grapevine Cultivar Tannat Berries Is Conferred Primarily by Genes That Are Not Shared with the Reference Genome. *The Plant Cell*, 25(12), pp.4777–4788.

De Simoni Gouveia, J.J. et al., 2014. Identification of selection signatures in livestock species. *Genetics and molecular biology*, 37(2), pp.330–42.

Smulders, M.J.M. et al., 2008. Structure of the genetic diversity in black poplar (Populus nigra L.) populations across European river systems: Consequences for conservation and restoration. *Forest Ecology and Management*, 255(5–6), pp.1388–1399.

Stebbins, G.L., 1958. The inviability, weakness, and sterility of interspecific hybrids. *Advances in genetics*, 9, pp.147–215.

Stölting, K.N. et al., 2015. Genome-wide patterns of differentiation and spatially varying

selection between postglacial recolonization lineages of Populus alba (Salicaceae), a widespread forest tree. *The New phytologist*, 207(3), pp.723–34.

Studer, A. et al., 2011. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nature genetics*, 43(11), pp.1160–3.

Sun, L. et al., 2014. Identification and analysis of genome-wide SNPs provide insight into signatures of selection and domestication in channel catfish (Ictalurus punctatus). *PLoS ONE*, 9(10), pp.1–10.

Sveinsson, S. et al., 2013. Transposon fingerprinting using low coverage whole genome shotgun sequencing in cacao (Theobroma cacao L.) and related species. *BMC genomics*, 14, p.502.

TAGI, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814), pp.796–815.

Tautz, D., 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic acids research*, 17(16), pp.6463–71.

Tellone, E. et al., 2015. Resveratrol: A Focus on Several Neurodegenerative Diseases. *Oxidative Medicine and Cellular Longevity*, 2015, pp.1–14.

Tetreault, H.M. & Ungerer, M.C., 2016. Long Terminal Repeat Retrotransposon Content in Eight Diploid Sunflower Species Inferred from Next-Generation Sequence Data. *G3: Genes, Genomes, Genetics*, 6(8), pp.2299–2308.

This, P. et al., 2004. Development of a standard set of microsatellite reference alleles for identification of grape cultivars. *Theoretical and Applied Genetics*, 109, pp.1448–1458.

This, P. et al., 2007. Wine grape (Vitis vinifera L.) color associates with allelic variation in the domestication gene VvmybA1. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 114(4), pp.723–30.

This, P., Lacombe, T. & Thomas, M.R., 2006. Historical origins and genetic diversity of wine grapes. *Trends in genetics : TIG*, 22(9), pp.511–9.

Tuskan, G. a. et al., 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, 313(5793), p.1596.

VanBuren, R. et al., 2015. Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature*, 527(7579), pp.508–511.

Vavilov, N.I., 1938. Plant resources of the world and their utilization for plant breeding. *Mathematics and Natural Science in the U.S.S.R.: Essays on the development of mathematics and natural science during the past 20 years*, pp.575–595.

Di Vecchi-Staraz, M. et al., 2009. Low level of pollen-mediated gene flow from cultivated to

wild grapevine: consequences for the evolution of the endangered subspecies Vitis vinifera L. subsp. silvestris. *The Journal of heredity*, 100(1), pp.66–75.

Velasco, R. et al., 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PloS one*, 2(12), p.e1326.

Velasco, R. et al., 2010. The genome of the domesticated apple (Malus × domestica Borkh.). *Nature Genetics*, 42(10), pp.833–839.

Vezzulli, S. et al., 2008. A SNP transferability survey within the genus Vitis. *BMC Plant Biology*, 8, p.128.

Vincent, H. et al., 2013. A prioritized crop wild relative inventory to help underpin global food security. *Biological Conservation*, 167, pp.265–275.

Vitti, J.J., Grossman, S.R. & Sabeti, P.C., 2013. Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, 47(1), pp.97–120.

Voight, B.F. et al., 2006. A Map of Recent Positive Selection in the Human Genome. *PLoS Biology*, 4(3), p.e72.

Walker, A.R. et al., 2007. White grapes arose through the mutation of two similar and adjacent regulatory genes. *The Plant journal : for cell and molecular biology*, 49(5), pp.772–85.

Wall, J.D., Andolfatto, P. & Przeworski, M., 2002. Testing models of selection and demography in Drosophila simulans. *Genetics*, 162(1), pp.203–216.

Watson, J.D. et al., 2008. The Genetic Code. In *Molecular Biology of the Gene*. San Francisco: Pearson Education, Inc., publishing as Benjamin Cummings, p. 522.

Weber, E., 2012. *Das kleine Buch der botanischen Wunder*, C.H.Beck.

Weber, J.L. & Myers, E.W., 1997. Human whole-genome shotgun sequencing. *Genome Res.*, 7, pp.401–409.

Wenke, T. et al., 2011. Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *The Plant cell*, 23(9), pp.3117–28.

Xu, J. & Kasha, K.J., 1992. Transfer of a dominant gene for powdery mildew resistance and DNA from Hordeum bulbosum into cultivated barley (H. vulgate). *Theoretical and Applied Genetics*, 84, pp.771–777.

Yan, K. et al., 2012. Stress-Induced Alternative Splicing Provides a Mechanism for the Regulation of MicroRNA Processing in Arabidopsis thaliana. *Molecular Cell*, 48(4), pp.521–531.

Yang, Z., 2002. Inference of selection from multiple species alignments. *Current Opinion in Genetics & Development*, 12(6), pp.688–694.

Yuan, Y., SanMiguel, P.J. & Bennetzen, J.L., 2003. High-Cot sequence analysis of the maize genome. *Plant Journal*, 34, pp.249–255.

Zakharenko, L.P., Kovalenko, L. V & Mai, S., 2007. Fluorescence in situ hybridization analysis of hobo, mdg1 and Dm412 transposable elements reveals genomic instability following the Drosophila melanogaster genome sequencing. *Heredity*, 99(5), pp.525–530.

Zecca, G. et al., 2012. The timing and the mode of evolution of wild grapes (Vitis). *Molecular phylogenetics and evolution*, 62(2), pp.736–47.

Zecca, G. et al., 2010. Wild grapevine: silvestris, hybrids or cultivars that escaped from vineyards? Molecular evidence in Sardinia. *Plant biology (Stuttgart, Germany)*, 12(3), pp.558–62.

Zeven, A., Knott, D. & Johnson, R., 1983. Investigation of linkage drag in near isogenic lines of wheat by testing for seedling reaction to races of stem rust, leaf rust and yellow rust. *Euphytica*, 32(1983), pp.319–327.

Zhang, X. et al., 2006. Identification of a drought tolerant introgression line derived from Dongxiang common wild rice (O. rufipogon Griff.). *Plant Molecular Biology*, 62(1–2), pp.247–259.

Zheng, S.-X., Xiao, S. & Chye, M.-L., 2012. The gene encoding Arabidopsis acyl-CoA-binding protein 3 is pathogen inducible and subject to circadian regulation. *Journal of Experimental Botany*, 63(8), pp.2985–3000.

Zhou, H. et al., 2005. Anticancer activity of resveratrol on implanted human primary gastric carcinoma cells in nude mice. *World J Gastroenterol*, 11(2), pp.280–284.

Zhou, Z. et al., 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology*, 33(4), pp.408–414.

Zohary, D., Hopf, M. & Weiss, E., 2012. Grapevine: Vitis vinifera. In *Domestication of Plants in the Old World: The Origin and Spread of Cultivated Plants in West Asia, Europe and the Nile Valley*. New York: Oxford University Press, pp. 121–126.

Zytnicki, M., Akhunov, E. & Quesneville, H., 2014. Tedna: a transposable element de novo assembler. *Bioinformatics (Oxford, England)*, 30(18), pp.2656–8.

Danksagung

Eidesstattliche Erklärung

Curriculum vitae