LIB Leibniz-Institut zur Analyse des Biodiversitätswandels

HAW HAMBURG

Hochschule für Angewandte Wissenschaften Hamburg

Fakultät Life Science

Herstellung und Analyse einer ddRAD - Bibliothek von *Gampsocleis glabra*

Bachelorarbeit

im Studiengang Biotechnologie

Vorgelegt von Carsten Bruns

[redacted]

Abgabe: 12.12.2022

Erster Gutachter: Prof. Dr. Julien Béthune HAW Hamburg

Zweiter Gutachter: Dr. Oliver Hawlitschek LIB Hamburg

Die Bachelorarbeit wurde betreut und erstellt im Molekularlabor des Leibnitz-Institut zur Analyse der Biodiversitätswandels (LIB) in Hamburg

Inhaltsverzeichnis

Zusammenfassung

In dieser Bachelorarbeit wurde eine ddRAD-Bibliothek (double digest reactions-site associated DNA) hergestellt, um die Heideschrecke *Gampsocleis glabra* populationsgenetisch zu analysieren. Das Ziel ist es, eine Aussage zu treffen, ob sich der Genotyp, der isoliert vorkommenden Heideschrecke verändert hat, so dass sich neue Populationen gebildet haben. Dafür wurden Proben aus Deutschland (Klietz, Munster, Rheinmetall), Niederlande, Slowakei und Ungarn untersucht. Das Protokoll für die Herstellung der ddRAD-Bibliothek wurde neu zusammengestellt, deswegen ist ein weiteres Ziel dieser Arbeit das Protokoll qualitativ zu untersuchen, ob es für die Analyse in der Populationsgenetik eingesetzt werden kann und um es weiter zu verbessern.

Eine ddRAD-Bibliothek ist eine DNA-Bibliothek, die es ermöglicht, mehrere DNA-Fragmente von Proben, die durch eine Digestion mit zwei Restriktionsenzymen entstehen, mit unterscheidbarem Adapter und Primer (Barcode, Index) und einer NGS-Sequenzierung populationsgenetisch zu untersuchen. Es wird eine ddRAD-Bibliothek eingesetzt, weil sie kostengünstiger, effizienter und mehr genetische Informationen bzw. Polymorphismen erzeugt als anderen Methoden z.B. „SNP-Chips“.

Für die Beantwortung der Fragen wurden zwei ddRad-Bibliotheken hergestellt, jeweils mit 24 Proben aus unterschiedlichen Fundorten. Für die Herstellung wurde die Restriktionsenzyme *SbfI* und *MseI* für die Digestion eingesetzt. Die Unterscheidung der Proben wurde mit den spezifischen Adaptern und Primern für die Illumina-NGS-Sequenzierung gewährleistet. Die Fragmentierung erfolgte mit magnetischen Beads und BluePippin. Jeder Schritt der Herstellung wurde mit dem Elektrophorese-System Tapestation kontrolliert. Die Sequenzierung wurde mit eine Illumina MiSeq durchgeführt. Die bioinformatische Bearbeitung und Vorbereitung der Reads erfolgte mit den Programmen STACKS und CUTADAPT. Anschließend wurde mit STRUCTURE und STRUCTURE HARVESTER die Populationsstruktur bestimmt. Dabei wurden jeweils die beiden einzelnen Bibliotheken und eine Zusammenführung der Daten der beiden Bibliothek analysiert. Die Analyse der Qualität wurde mit dem Programm FASTQC und MULTIQC durchgeführt. Die Einstellungen des Programms wurde so konzipiert, dass sie eine Schnellanalyse möglich macht.

Die beiden Bibliotheken sind nach dem Phred-Qualität-Wert, der für die meisten Sequenzen zwischen 30 und 35 liegt, für die Populationsgenetik einsetzbar. Die einzelnen Bibliotheken

haben Anteile von drei Populationen, die zusammengeführten Daten habe Anteile von vier Populationen im Genotyp. Es hat sich gezeigt, dass die Aussagen der drei Populationsstrukturen nicht einheitlich sind und durch die stochastische Bearbeitung (MCMC, Ad-hoc-Statistik) nicht vergleichbar sind. Für eine bessere Aussage über eine Populationsstruktur müsste die ddRAD-Bibliothek um mehrere Proben aus verschiedenen Fundorten vergrößert werden, die pro Ort die gleiche Anzahl haben.

Abkürzungsverzeichnis

Abkürzung	Bedeutung
ddRAD	Engl. double digest reactions-site associated DNA
PCR	Engl. Polymerase chain reaction
DNA	Desoxyribonukleinsäure
SbfI	Bezeichnung eines Restriktionenzym
MseI	Bezeichnung eines Restriktionenzym
NGS	Engl. Next Generation Sequencing
Tris	Tris(hydroxymethyl)aminomethan
NaCl	Natriumchlorid
EDTA	Ethylendiamintetraessigsäure
OH	Oliver Hawlitschek
MH	Martin Husemann
MCMC	Engl. Markov Chain Monte Carlo
K	Population
dNTPs	Deoxynucleodiftriphosphate
bp	Basenpaare
NEB	New England Biolabs
A	Adenin
T	Thymin
C	Cytosin
G	Guanin
SVK	Slowakei
HUN	Ungarn
NDL	Niederlande
DEUK	Deutschland, Klietz
DEUM	Deutschland, Munster
DEUR	Deutschland, Rheinmetall
kA	Keine Angaben
SNP	Engl. Single nucleotide polymorphism

Einleitung

Zielführung der Bachelorarbeit

Das Ziel dieser Bachelorarbeit ist die Herstellung einer ddRAD-Bibliothek mit einem neu zusammengestellten Protokoll und einer populationsgenetischen Analyse der Heideschrecke, *Gampsocleis glabra*. Die Heideschrecken werden untersucht, da sie in Deutschland nur selten und in isolierten Gebieten vorkommt z. B. auf militärischen Truppenübungsplätzen. Deswegen wird in dieser Arbeit untersucht, ob sich durch das isolierte Vorkommen der Genotyp verändert hat, so dass sich neue Populationsstrukturen ausgebildet haben. Da es sich um ein neues Protokoll handelt, ist ein weiteres Ziel, eine Aussage zu treffen, ob das Protokoll für Populationsgenetik eingesetzt werden kann und wie es verbessert werden kann.

Molekularbiologischer Hintergrund einer ddRAD-Bibliothek

In der Populationsgenetik werden DNA-Bibliotheken eingesetzt, um mehrere oder unterschiedliche Arten von Organismen gleichzeitig mit einer NGS-Sequenzierung bioinformatisch zu untersuchen. Ein weiterer Grund ist, dass wesentlich mehr Informationen über Polymorphismen generiert werden als bei alten Verfahren, z.B. mit Mikrosatelliten, mit diesen könnten nur bis zu 10 Polymorphismen untersucht werden. Bei einer NGS mit einer DNA-Bibliothek liegt die Zahl über 1000. Des Weiteren ist es eine kostengünstigere Methode in Gegensatz zur eine kompletten Genom-Untersuchung, weil nur Teilabschnitte untersucht werden.

Die Abkürzung NGS steht in englisch für ´´Next Generation Sequencing``. In Deutsch wird NGS als Hochdurchsatz-Sequenzierung bezeichnet.

Eine NGS-Sequenzierung wird dabei eingesetzt, weil die Sequenzierung wesentlich schneller als bei den anderen Verfahren abläuft, z.B. dem Sanger-Verfahren. Ein weiterer Grund ist, dass die DNA nur in der Länge von 100 bp bis 300 bp sequenziert wird. Dabei werden die Fragmente gleichzeitig sequenziert. [Reinard, 2021]. Diese Eigenschaft der NGS wird für die DNA-Bibliothek eingesetzt. Bei der Sequenzierung der Bibliotheken wird eine paired-end - Sequenzierung angewendet. Dabei werden die Sequenzen zweimal durchlaufen, einmal vorwärts (single-end, R1) von 5´- Ende bis 3´- Ende und ein zweites Mal rückwärts (paired-end, R2).

Allgemein ist eine DNA-Bibliothek ein DNA-Fragment und ein Adapter mit einem Barcode, der das Fragment einer Probe zuordnet. Bei der eingesetzte DNA-Bibliothek handelt es sich um eine ddRAD-Bibliothek. Die Bezeichnung ddRAD lautet in englischen „**d**ouble **d**igest **R**eactions-site **A**ssociated **D**NA“.

Eine ddRAD-Bibliothek wird eingesetzt, um die Genotypen effizient, genauer, kostengünstiger und ohne Vorwissen bzw. ohne eine Referenzgenom zu analysieren. Die anderen Methoden, z.B. „SNP-Chips“, die auf eine Microarray-Technologie basiert, ist zwar kostengünstig, aber dieses Verfahren benötigt das Wissen über die genomische Sequenz und deren Variabilität und sie gilt nur für eine Nukleotidstellen. Also eine Populationsanalyse von unbekannten Organismen bzw. Lebewesen, deren Genomsequenz noch nicht analysiert wurde, kann mit den SNP-Chips nicht untersucht werden, bzw. nur mit erheblichem Zeitaufwand. [Peterson, 2012]

Dadurch, dass die Heideschrecke selten in Deutschland vorkommt, gibt es keine Informationen über die Genomsequenzen, um andere Methoden (SNP-Chips) einzusetzen. Deswegen wird für die Populationsanalyse ein ddRAD-Bibliothek eingesetzt.

Ein Beispiel für den Einsatz einer ddRAD-Bibiliothek ist die Populationsanalyse der Baumwanze, *Halyomorpha halys*, die in Ostasien vorkommt, die sich über die weltweiten Handelswege verteilt hat und Schäden in der Landwirtschaft verursacht. Bei der Untersuchung hat sich gezeigt, dass sich fünf Varianten ausgebildet haben, jeweils zwei aus den Heimatorten in China und Japan und drei habe sich in befallenen Gebieten, z.B. USA und Europa, gebildet. [Yan, 2012]

Ein weiteres Beispiel für die Anwendung der ddRAd-Bibliothek ist in der Fischzucht, um neu entstehende Arten zu Klassifizierung bzw. populationsgenetisch zu untersuchen z.B. der Seesaibling, *Salvelinus alpinus*. Diese neue Art hat in der Zucht nicht das Produktionsvolumen, um diese mit teuren Methoden untersuchen zulassen, deswegen werden in der Zucht von neuen Arten kostengünstigere Methoden eingesetzt. [Pappas, 2021]

Der Aufbau einer ddRAD-Bibliothek ist in der Abbildung 1 dargestellt. In dieser Bibliothek wird die extrahierte DNA mit zwei Restriktionsenzyme verdaut, um die DNA in mehrere kleineren Fragmente zu schneiden, um Reaktionsstellen für die Adapter P1 und P2 zu erhalten.

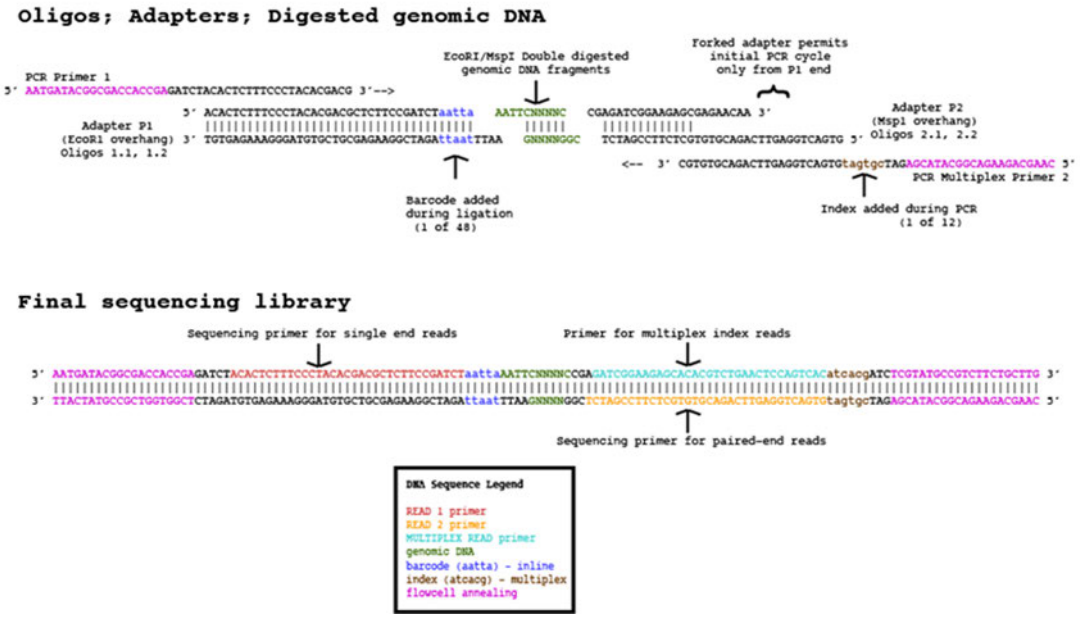

Abbildung 1: Schema eine ddRAD-Bibliothek.
Die Abbildung zeigt die Elemente einer ddRAD-Bibliothek an. Grün ist das DNA-Fragment, Blau der Barcode, Braun ist der Index-Code der Primer. Rot und gelb geben die Anlagerungsstellen der Sequenzierprimer für die single-end Reads (R1) und die für die paired-end Reads (R2) an. [Peterson, ddRAD-Protocol]

Durch eine Ligation werden die Adapter P1 und P2 an die Reaktionsstellen ligiert. Der Adapter P1 enthält den Barcode, der das Fragment einer Probe zuordnet. Der Aufbau des Adapter P1 ist in Tabelle 1 aufgeführt. Bei den Adaptern handelt es sich um kurzkettige Oligonukleotide. Die vier Basen in Grün sind die Reaktionsstellen des Restriktionsenzym von *SbfI* 5´-TGCA - 3´ und von *MseI* 5´- TAA - 3´. Die Roten Basen sind die Barcode 5´ - GCATG - 3´. Die blauen Basen sind Abstandhalter zum Barcode und unterstützt die Ligase bei der Anlagerung an die Reaktionsstelle. Jede Probe der Bibliothek bekommt einen anderen P1 Adapter bzw. Barcode zugewiesen.

Tabelle 1: Adapteraufbau.
Rot sind die Basen des Barcodes. Die grün eingefärbten Basen sind die Reaktionsstellen der Enzyme *SbfI* in P1 und *MseI* in P2. Die blauen Basen sind die Abstandshalter.

Adapter	Sequenz
P1	5´- ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCATGCCTGCA - 3´
P2	5´ - TAAGATCGGAAGAGCGAGAACAA - 3´

Die beiden Adapter dienen für die PCR als Anlagerungsstellen für die beiden Primer. Der Rückwärts-Primer enthält einen Index, der eine weitere Unterscheidung der Fragmente ermöglicht. Durch den Barcode der Adapter und den Index der Primer ist es möglich, mehrere Proben in einem NGS-Lauf zu sequenzieren. [Peterson, 2012]

Eingesetzte Software für die Analyse einer ddRAD-Bibliothek

Für die Analyse der ddRAD-Bibliotheken wird das Programm STACKS benutzt. In der Qualitätsanalyse kommen die Programme FASTQC und MULTIQC und für die populationsgenetische Analyse das Programm STRUCTURE und das Web-Programm STRUCTURE HARVESTER zum Einsatz.

STACKS

Das Programm STACKS wurde von Julian Catchen entwickelt. Es wurde entwickelt, um die kurzen Sequenzen, die aus der NGS-Sequenzierung stammen, bioinformatisch zu verarbeiten. Die kurzen Sequenzen werden als Reads bezeichnet. Die Reads werden durch STACKS geordnet und anschließend werden daraus Allele und Genotypen mit einer maximalen Wahrscheinlichkeitsmethode gebildet, die wiederum mit den Individuen einer Population verglichen werden. Um die Daten zu verarbeiten, besteht STACKS aus mehreren Programmen, die nacheinander abgearbeitet werden. [Catchen, 2013]

In der Abbildung 2 wird die Programmstruktur von STACKS dargestellt. Das Unterprogramm *process_radtags* fasst die Rohdaten mit der Hilfe einer Barcode-Liste zusammen. Weiterhin kann das Programm die Reads reinigen, z.B. von Sequenzen mit schlechter Qualität, Adapterkontaminationen oder die Reads auf eine Länge zuschneiden. Das zweite Programm ist *ustacks,* entwickelt aus den Daten der einzelnen Arten de novo die Loci. Dafür wird ein K-mer Algorithmus eingesetzt, um die Polymorphismen zu identifizieren. Das Programm *cstacks* fasst die Loci der einzelnen Proben in einem Katalog zusammen. Mit dem *sstacks* Programm werde die Loci im Katalog mit den Loci die einzelnen Proben verglichen, um Allele zu bilden. [Catchen, 2011,2013]. Das *tsv2bam*-Programm strukturiert die Daten von *sstacks* in das Format *bam*, dabei werden die Daten von den paired-end-Reads mit integriert. Das Programm *gstacks* analysiert die Daten nach den Loci, also das Programm sucht in den einzelnen Proben nach diesen Loci. Es werden Contigs (Sequenzen, die sich überlappen und ein genomischen Bereich bilden. [www.genome.gov/genetics-glossary/Contig]) vom den single-end und den paired-end-Daten erzeugt und zusammengeführt. [Rochette, 2019]. Das letzte Programm *populations* filtert die Daten und schreibt den Datensatz in andere Formate um.

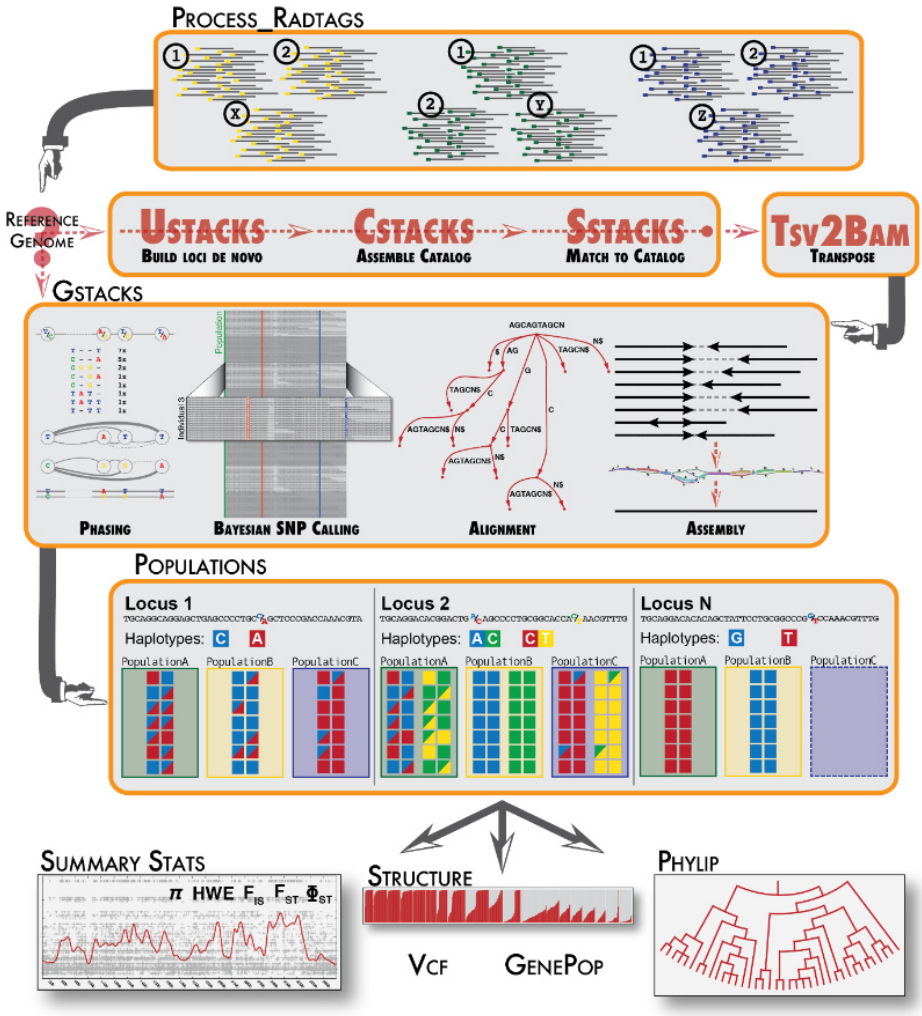

Abbildung 2: Programmstruktur von Stacks.
Das Schema zeigt die Bearbeitungsreihenfolge der STACKS-Unterprogramme an. Process_radtags sortiert die Sequenzen. Die Programme *ustack*, *cstack* und *sstack* bilden die Loci, katalogisieren die Locis und vergleichen die unterschiedlichen Kataloge der Proben miteinander. *Tsv2bam* schreibt die Daten in das *bam*-Format um. *gstacks* identifiziert SNPs und bildet haploide Genotype pro Probe aus. Das Programme *populations* bestimmt die populationsgenetischen Statistiken und herstellt Formate z.B. für STRUCTURE. [Catchen, STACKS-Manual]

FASTQC

Für die Einschätzung der Qualität der ddRAD-Bibliothek wird das Programm FASTQC von Simon Andrew von Babraham bioinformatics Institut genutzt. Das Programm nutzt unter anderen die Phred-Qualitäts-Punktzahl für eine qualitative Einschätzung der DNA-Bibliothek. Die Phred-Punktzahl steht logarithmisch mit der Fehlerwahrscheinlichkeit der Basen bei einer Sequenzierung in Beziehung.

$$q = -10 \cdot \log_{10} p \quad oder \quad p = 10^{\frac{-q}{10}}$$

Wenn die Base z.B. einen Wert von 20 hat, liegt die p bei 1/100 bzw. bei 100 Base wird nur eine Base falsch dargestellt. [Ewing, 1998] Die Werte für die Phred-Punktzahl steht in dem Fastq-Format als ASCII-Code unter den Basensequenzen.

FASTQC erhebt noch zwei weitere Parameter, dies sind der Anteil an PCR-Duplikationen und der Anteil der überrepräsentierten Sequenzen. Die PCR-Duplikationen entstehen durch die PCR, indem Kopie von der ursprünglichen DNA amplifiziert werden und damit den Anteil, der der daraus resultierte Reads erhöht. Überrepräsentierten Sequenzen sind Sequenzen, die häufiger in Reads vorkommen als erwartet, z.B. sind das Adaptersequenzen oder andere Kontaminationen. Es kann aber auch ein Anzeichen dafür sein, dass die biologische Diversität gering ist.

STRUCTURE und STRUCTURE HARVESTER

STRUCTURE basiert auf einem Modell von Pritchard et al., das für Populationsuntersuchungen eingesetzt wird. Das Modell beruht auf eine Bayesianische Stochastik die mit der Anwendung eines Markov-Chain-Monte-Carlo-Algorithmus (MCMC) eine Aussage über eine Populationsstruktur treffen kann. Eine weiter Annahme für das Modell ist das Hardy-Weinberg-Gleichgewicht. Dieses Gleichgewicht definiert eine ideale Population, in der es keine genotypischen Veränderungen gibt. [Pritchard, 2000]

Die Bayesianische Stochastik ermöglicht, das aus den Daten *D*, einer Hypothese *K* und der maximalen Wahrscheinlichkeit $P(K|D)$ aus den beiden, eine neue Aussage über die Hypothese getroffen werden kann.

$$P(K_n|D) = \frac{P(D|K) \cdot P(K)}{\sum_i P(D|K_i) \cdot \ P(K_i)}$$

K ist in dieser Arbeit die Anzahl der Populationen, die durch die Fundorte vorgegeben wurde. *D* sind die Sequenzen, die aus der Sequenzierung einer ddRAD-Bibliothek stammen. $P(D|K)$ Ist die maximale Wahrscheinlichkeit, die durch das Programm STACKS erzeugt wird bzw. die Allelehäufigkeit, die aus den bearbeiteten Sequenzen resultiert. K_n ist die neu bestimmte Wahrscheinlichkeit über *K*. Die Summe im Nenner summiert alle möglichen K_i auf. Da diese Summe sehr viel Rechnerleistung benötigt, wird der MCMC-Algorithmus eingesetzt. [Knoop, 2009]

Der MCMC-Algorithmus nimmt nur Stichproben durch eine Markow-Kette aus der Wahrscheinlichkeitsberechnung. Die Markow-Kette ist gedächtnislos, d.h. die zukünftigen Zustände sind nur vom aktuellen Zustand abhängig und nicht von vergangenen Zuständen. Wenn die Berechnung lang genug läuft, nähert sich die Kette einem stationären Bereich an, der als eine Wahrscheinlichkeitsverteilung aufgefasst werden kann. In diesen Bereich hängt die Verteilung nicht mehr von den Anfangsbedingungen ab. Aus dieser stationären Verteilung wird eine Stichprobe genommen und diese entspricht dann die Wahrscheinlichkeit $P(K_n|D)$. [Knoop, 2009]

Das Web-Programm STRUCTURE HARVESTER wird genutzt, um ein stabiles ΔK zu bestimmen. Das Programm STRUCTURE HARVESTER benutzt, um ΔK zu bestimmen, die Evanno-Methode. Die Methode wird angewendet, weil die Log-Wahrscheinlichkeitsverteilung, die STRUCTURE bildet, keine genaue Aussage über ein stabiles K liefert bzw. nicht die dominierende Anzahl der Population K anzeigt, die durch die Allelehäufgkeit aus den Proben entsteht. Die Evanno-Methode hat gezeigt, dass ein stabiles K mit einer Ad-Hoc-Statistik und durch eine Änderungsrate zweiter Ordnung zwischen den Daten von verschiedenen K‘s ein wahres K bestimmt werden kann. [Evanno, 2005]

$$L'(K) = L(K) - L(K-1)$$

$$L''(K) = L'(K+1) - L'(K)$$

$$\Delta K = \frac{L''(K)}{s(L(K))}$$

Dafür wird eine Änderungsrate der Wahrscheinlichkeitsverteilung zweiter Ordnung L‘‘(K) gebildet und die durch die Standardabweichung der Log-Wahrscheinlichkeit L(K) geteilt.

Eine Fragestellung für die Populationsanalyse von *Gampsocleis glabra*

Da die Proben aus 6 verschiedenen Standorten kommen und die Heideschrecke isoliert vorkommt, wird vermutet, dass die Population der Heideschrecke sich zwischen den Standorten unterscheidet und sich schon genotypische Unterschiede ausgebildet haben. Im maximalen Fall kann es 6 Population der Heideschrecke geben, jeweils eine pro Standort.

Material und Methoden

Liste der eingesetzten Materialen, Geräte und Programme

In den Tabelle 2 bis 4 sind alle Geräte, Chemikalien und Programme aufgeführt, die bei der Bearbeitung der Aufgaben genutzt wurden.

Tabelle 2: Liste der eingesetzten Geräte

Geräte	Name	Firma
Thermocycler	Biometra Tone	analytikjena
Thermoblock	Thermo Shaker TS-100C	biosan
Zentrifuge	Sprout	Biozym
Elektrophorese-System	4200 Tapestation	Agilent Technologies
NGS-Sequenzer	Illumina MiSeq	Illumina
Photometer	Nanotrop 2000	Thermo scientific
Fragmentierer	Bluepippin	sage science
Vortexer	Vortex-Genie 2	Scientific Industries

Tabelle 3: Liste der benutzten Materialien

Materialien	Firma
SbfI	NEB
MseI	NEB
Cutsmart Puffer 10x	NEB
GeneJet PCR Purification Kit	Thermo Scientific
Magnetische Beads MagSi-NGS Plus	MagnaMedics
Oligonucleotide P1 und P2 für die Adapter	Eurofins Genomics
Annealing Puffer 10x: 100 mM Tris 10 mM EDTA 500 mM NaCl	 Roth Roth VWR chemicals
Ligase T4	NEB
Primer ILLPCR 1	Eurofins Genomics
Primer ILLPCR2_01	Eurofins Genomics
Primer ILLPCR2_02	Eurofins Genomics
PHUSION buffer HF 5x	NEB
dNTPs	Jena Bioscience
PHUSION Polymerase	NEB
Screentape HS D1000, D1000, D5000	Agilent Technologies
Pippin Gel Cassette Internal Standard 250 bp – 1,5 kp 1,5 % Agarose, Dye-Free, Marker R2	sage science
BluePippin Reagent Kit Internal Standard Mix 250 bp -1.5 kb Marker R2	sage science

Tabelle 4: Liste der genutzten Programme

Programme	Version
STACKS	2.60 [Catchen]
FASTQC	0.11.9 [Andrews]
MULTIQC	1.12
STRUCTURE	2.3.4 [Pritchard et. Al]
STRUCTURE HARVESTER	Web v0.6.94 July 2014, Plot vA.1 November 2012, Core vA.2 July 2014 [Earl]
CUTADAPT	3.4 [Rahmann]

Information zur *Gampsocleis glabra*

Das Insekt *Gampsocleis glabra* wird auf Deutsch als Heideschrecke bezeichnet. Die Schrecke, ist von Frankreich bis in die Mongolei lokal verbreitet. In Deutschland ist *G. glabra* vom Aussterben bedroht und ist überwiegend in der Lüneburger Heide oder in Klietzer Heide zu finden. Generell ist die Heideschrecke in trockenwarmen Grassteppen, Heiden oder auf alten Truppenübungsplätzen zu finden. Die Färbung der Heideschrecke variieren zwischen einen grünen, gelben oder grauen Farbton. Die Flügel sind länger als der Hinterteil, aber nicht länger als das Knie der Hinterbeine. Ein besonderes Merkmal bei den Weibchen ist das körperlange Legerohr. [http://www.orthoptera.ch/arten/item/gampsocleis-glabra]

Für die ddRAD-Bibliothek wurden *G. glabra* Proben von verschieden Orten eingesetzt. Die Proben kommen aus Munster, Klietz und Rheinmetall in Deutschland, aus Slowakei, Ungarn und aus den Niederlanden. [Tab. 5, 6]

In der Tabelle 5 sind die Proben aufgezählt, die für die erste ddRAD-Bibiliothek ML1 eingesetzt wurden und in der Tabelle 6 sind die Proben für die ddRAD-Bibliothek ML2 verzeichnet. In den Tabellen ist die Proben-ID, die Gattungs- und Arten-Bezeichnung, das Geschlecht, das Funddatum, der Fundort und die Abkürzung der Sammler aufgeführt.

Tabelle 5: Proben der Ersten ddRAd-Bibliothek ML1.

Die Abkürzungen: m steht für männlich, f für weiblich, k.A.: keine Angaben, SVK: Slowakei, HUN: Ungarn, DEU: Deutschland, NLD: Niederlande. Die Sammler sind Oliver Hawlitschek (OH) und Martin Husemann (MH)

Proben-ID	Gattung	Art	Geschlecht	Funddatum	Fundort	Sammler
ML023	*Gampsocleis*	*glabra*	m	k. A.	SVK	OH
ML025	*Gampsocleis*	*glabra*	m	k. A.	SVK	OH
ML029	*Gampsocleis*	*glabra*	f	k. A.	SVK	OH
ML030	*Gampsocleis*	*glabra*	m	k. A.	SVK	OH
ML032	*Gampsocleis*	*glabra*	m	k. A.	HUN	OH
ML033	*Gampsocleis*	*glabra*	m	k. A.	HUN	OH
ML034	*Gampsocleis*	*glabra*	f	k. A.	HUN	OH
ML035	*Gampsocleis*	*glabra*	m	k. A.	HUN	OH
ML050	*Gampsocleis*	*glabra*	m	17.08.2020	DEU, Klietz	MH
ML051	*Gampsocleis*	*glabra*	m	17.08.2020	DEU, Klietz	MH
ML061	*Gampsocleis*	*glabra*	m	17.08.2020	DEU, Klietz	MH
ML073	*Gampsocleis*	*glabra*	f	28.08.2020	DEU, Rheinmetall	MH
ML088	*Gampsocleis*	*glabra*	k. A.	31.07.2020	NLD, Oldebrack	k. A.
ML089	*Gampsocleis*	*glabra*	k. A.	31.07.2020	NLD, Oldebrack	k. A.
ML090	*Gampsocleis*	*glabra*	k. A.	31.07.2020	NLD, Oldebrack	k. A.
ML091	*Gampsocleis*	*glabra*	k. A.	31.07.2020	NLD, Oldebrack	k. A.
ML098	*Gampsocleis*	*glabra*	k. A.	31.07.2020	NLD, Oldebrack	k. A.
ML102	*Gampsocleis*	*glabra*	f	k. A.	k. A.	OH
ML104	*Gampsocleis*	*glabra*	m	k. A.	DEU, Munster	MH
ML107	*Gampsocleis*	*glabra*	m	k. A.	DEU, Munster	MH
ML110	*Gampsocleis*	*glabra*	m	k. A.	DEU, Rheinmetall	MH
ML111	*Gampsocleis*	*glabra*	m	k. A.	DEU, Rheinmetall	MH
ML112	*Gampsocleis*	*glabra*	m	k. A.	DEU, Rheinmetall	MH
ML114	*Gampsocleis*	*glabra*	m	k. A.	DEU, Rheinmetall	MH

Tabelle 6: Proben der zweiten ddRAD-Bibliothek ML2

Die Abkürzung m steht für männlich, f: weiblich, k.A.: keine Angaben, SVK: Slowakei, HUN: Ungarn, DEU: Deutschland, NLD:Niederlande. Die Sammler sind Oliver Hawlischek (OH) und Martin Husemann (MH)

Proben-ID	Gattung	Art	Geschlecht	Funddatum	Fundort	Sammler
ML020	*Gampsocleis*	*glabra*	f	k. A.	SVK	OH
ML021	*Gampsocleis*	*glabra*	f	k. A.	SVK	OH
ML026	*Gampsocleis*	*glabra*	f	k. A.	SVK	OH
ML027	*Gampsocleis*	*glabra*	f	k. A.	SVK	OH
ML036	*Gampsocleis*	*glabra*	f	k. A.	k. A.	OH
ML052	*Gampsocleis*	*glabra*	m	17.08.2020	DEU, Klietz	MH
ML054	*Gampsocleis*	*glabra*	m	17.08.2020	DEU, Klietz	MH
ML058	*Gampsocleis*	*glabra*	m	17.08.2020	DEU, Klietz	MH
ML062	*Gampsocleis*	*glabra*	m	17.08.2020	DEU, Klietz	MH
ML070	*Gampsocleis*	*glabra*	f	28.08.2020	DEU,Rheinmetall	MH
ML071	*Gampsocleis*	*glabra*	f	28.08.2020	DEU,Rheinmetall	MH
ML075	*Gampsocleis*	*glabra*	m	28.08.2020	DEU,Rheinmetall	MH
ML076	*Gampsocleis*	*glabra*	m	28.08.2020	DEU,Rheinmetall	MH
ML077	*Gampsocleis*	*glabra*	m	28.08.2020	DEU,Rheinmetall	MH
ML078	*Gampsocleis*	*glabra*	m	14.08.2020	DEU,Rheinmetall	MH
ML092	*Gampsocleis*	*glabra*	k. A.	31.07.2020	NLD, Oldebrack	k. A.
ML093	*Gampsocleis*	*glabra*	k. A.	31.07.2020	NLD, Oldebrack	k. A.
ML094	*Gampsocleis*	*glabra*	k. A.	31.07.2020	NLD, Oldebrack	k. A.
ML095	*Gampsocleis*	*glabra*	k. A.	31.07.2020	NLD, Oldebrack	k. A.
ML099	*Gampsocleis*	*glabra*	m	k. A.	k. A.	k. A.
ML100	*Gampsocleis*	*glabra*	f	k. A.	k. A.	OH
ML101	*Gampsocleis*	*glabra*	m	k. A.	k. A.	OH
ML109	*Gampsocleis*	*glabra*	m	k. A.	DEU, Rheinmetall	MH
ML115	*Gampsocleis*	*glabra*	m	k. A.	DEU, Rheinmetall	MH

Extraktion der Proben

Die DNA von den Heideschrecken-Proben wurden mit dem Protokoll „High Salt Extraction Protocol“ von Robert Paxtron von den Mitarbeitern des Labors isoliert.

Herstellung der ddRAD-Bibliothek

Für zwei ddRAD-Bibliotheken wurden jeweils 24 DNA-Proben von *G. glabra* eingesetzt. Die Methode setzt sich aus mehreren Teilen zusammen. In der Abbildung 3 ist die Herstellung der DNA-Bibliothek in einem Grundfließbild dargestellt. Für die erste Bibliothek ML1 wurde nach der Digestion, der Ligation für jede Probe und der PCR eine Konzentrationsmessung durchgeführt, um eine weiter Kontrolle um die Herstellung zu haben.

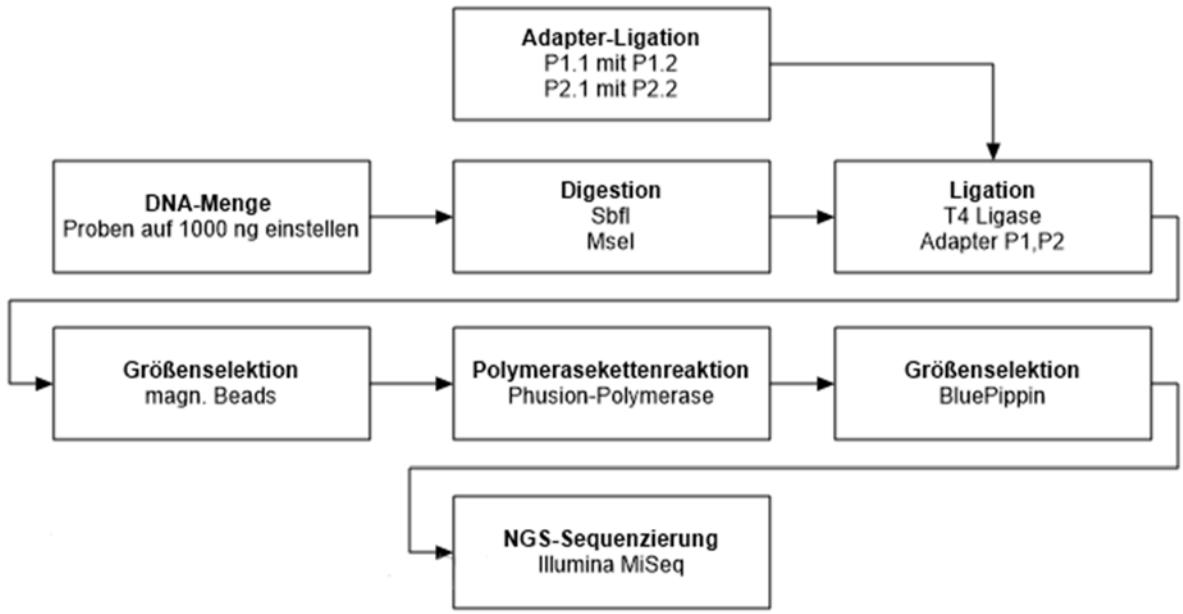

Abbildung 3: Grundfließbild für die Herstellung einer ddRAD-Bibliothek.
Das Fließbild zeigt die Reihenfolge, indem die ddRAD-Bibliothek bearbeitet worden ist. Als erste wurde die DNA-Menge eingestellt auf 1000 ng. Danach erfolgt die Digestion mit den Restriktionsenzymen *SbfI* und *MesI* parallel dazu wurden die Oligonukleotide der Adapter P1 und P2 ligiert. Die verdauten Proben werden durch die *Ligase T4* mit den Adaptern ligiert. Nur bei der Bibliothek ML2 erfolgt eine Größenselektion mit magn. Beads. Danach wurde eine PCR mit der *Phusion-Polymerase* durchgeführt. Vor der Sequenzierung wurde bei der Bibliothek eine Größenselektion mit den Fragmentier BluePippin durchgeführt. Anschließend wurde die ddRAD-Bibliothek durch eine NGS-Sequenzierung mit einem Illumina MiSeq sequenziert.

DNA - Menge

Als erstes wurde die DNA-Konzentration der Proben bestimmt, um die DNA-Menge von 1000 ng einzustellen. Für die Messung wurde ein Nanodrop eingesetzt. Die Einstellung der DNA-Menge ist wichtig, damit beim Zusammenführen der Proben die einzelnen Proben die gleiche Menge DNA haben, so dass bei der Analyse die Reads gleichmäßig sind und keine einzelne

Probe dominiert. Um sicherzustellen, dass das Volumen nicht zu groß zum Einstellen der Menge ist, sollte die DNA-Konzentration der Proben in einen Bereich von 50 ng/µL bis 100 ng/µL liegen.

Digestion

Um Anlagerungsstellen für die Adapter und die genomische DNA zu fragmentieren, wurde eine Digestion (ein Verdau) mit den Restriktionsenzymen *SbfI* und *MseI* durchgeführt. In den 50 µL Reaktionsansatz wurden 1000 ng DNA von den Proben pipettiert. Hinzu kamen jeweils 1 µL von den beiden Restriktionsenzyme und 5 µL CutSmart Puffer 10x. Das restliche Volumen wird mit ddH_2O aufgefüllt. Anschließend wurde der Ansatz in einem Thermocycler (Biometra) für 15 min bei 37 °C inkubiert und danach für 20 min bei 80 °C inaktiviert. Der Verdau richtet sich nach den Angaben von NEB.

Um die Enzyme zu entfernen und die Proben aufzureinigen, wurde die erste Bibliothek mit dem GeneJet PCR Purification Kit (ThermoScientific) und die zweite Bibliothek mit magnetischen Beads MagSi-NGS Plus aufgereinigt.

Nach der Digestion und Aufreinigung wurden die Proben mit der Tapestation kontrolliert, ob die Digestion funktioniert. Die Tapestation ist ein Elektrophorese-System, das für die Kontrolle der DNA-Probe eingesetzt wird. Die Tapestation setzt für die Messung verschieden Screentapes ein (D1000 oder HS D1000, D5000). Nach der Digestion wurde das Screentape D5000 eingesetzt, was DNA-Fragmente bis 10000 bp anzeigt. Dafür wurden aber nicht alle Proben kontrolliert, sondern nur ein paar Proben. Für die Durchführung der Tapestation Elektrophorese wurde mit dem Protokoll der Screentapes durchgeführt.

Adapter-Ligation

Die beiden Adapter P1 und P2 liegen als einzelsträngige Oligonukleotide vor, die erst mit einer Ligation durch Senkung der Temperatur von 97,5 °C bis 21 °C zusammengefügt werden müssen. Für die Ligation wurden für beide Adapter eine 40 µM Lösung hergestellt. Für die P2-Adapter wurden in 100 µL, jeweils 40 µL für die beiden Oligonukleotide P2.1 und P2.1, 10 µL des Annealing-Puffer 10x und 10 µL ddH_2O pipettiert. In der Tabelle 7 sind die beiden Oliognukleotide des Adapters P2 mit ihren Sequenzen in Vorwärts- (F) und Rückwärtsrichtung (R) aufgeführt.

Tabelle 7: Basensequenz des Adapter P2.
In der Tabelle sind die Basenreihenfolge der Oligonukleotide P2.1 und P2.2 dargestellt.

Name	Richtung	Adapter-Sequenzen
P2.1	F	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
P2.2	R	TAAGATCGGAAGAGCGAGAACAA

Da der P1-Adapter den Barcode enthält, mussten 24 verschiedene Lösungen angesetzt werden. Dafür wurden für je eine Probe in 10 µL Reaktionsansatz, 1 µL des Annealing Puffer 10x, 1 µL ddH_2O und je 4 µL für die beiden Oligonukleotide P1.1 und P1.2 pipettiert. Wegen der verschiedenen Barcodes mussten für jede Probe eine anderes Oligonukleotidepaar gewählt werden. Die Tabelle 8 zeigt die Adapterteile von P1 an. Es wird die Sequenz der einzelnen P1-Adapter dargestellt und zeigt separat den Barcode, der in den jeweiligen P1 Adapter enthalten ist, an.

Tabelle 8: Basensequenz des Adapter P1.

In dieser Tabelle sind die Zuordnungen der beiden Oligonukleotide von Adapter P1 dargestellt inklusive zu welchem P1-Adapter die einzelnen Barcodes zugeordnet sind.

Name	Richtung	Barcode	Adapter-Sequenz
P1.1 - 01	F	GCATG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCATGCCTGCA
P1.1 - 02	F	AACCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAACCACCTGCA
P1.1 - 03	F	CGATC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGATCCCTGCA
P1.1 - 04	F	TCGAT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGATCCTGCA
P1.1 - 05	F	TGCAT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCATCCTGCA
P1.1 - 06	F	CAACC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAACCCCTGCA
P1.1 - 07	F	GGTTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGTTGCCTGCA
P1.1 - 08	F	AAGGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGGACCTGCA
P1.1 - 09	F	ACACA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACACACCTGCA
P1.1 - 10	F	ATACG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATACGCCTGCA
P1.1 - 11	F	CTTGG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTGGCCTGCA
P1.1 - 12	F	GAGTC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAGTCCCTGCA
P1.1 - 13	F	AGCTA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCTACCTGCA
P1.1 - 14	F	ACTTC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTTCCCTGCA
P1.1 - 15	F	CATAT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCATATCCTGCA
P1.1 - 16	F	CGGCT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGGCTCCTGCA
P1.1 - 17	F	CTGAT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATCCTGCA
P1.1 - 18	F	GCTGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTGACCTGCA
P1.1 - 19	F	GTAGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTAGTCCTGCA
P1.1 - 20	F	GTCCG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCCGCCTGCA
P1.1 - 21	F	TATAC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTATACCCTGCA
P1.1 - 22	F	ATGAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATGAGCCTGCA
P1.1 - 23	F	TGGAA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGAACCTGCA
P1.1 - 24	F	TTACC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTACCCCTGCA
P1.2 - 01	R	CATGC	GGCATGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 02	R	TGGTT	GGTGGTTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 03	R	GATCG	GGGATCGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 04	R	ATCGA	GGATCGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 05	R	ATGCA	GGATGCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 06	R	GGTTG	GGGGTTGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 07	R	CAACC	GGCAACCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 08	R	TCCTT	GGTCCTTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 09	R	TGTGT	GGTGTGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 10	R	CGTAT	GGCGTATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 11	R	CCAAG	GGCCAAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 12	R	GACTC	GGGACTCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 13	R	TAGCT	GGTAGCTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 14	R	GAAGT	GGGAAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 15	R	ATATG	GGATATGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 16	R	AGCCG	GGAGCCGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 17	R	ATCAG	GGATCAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 18	R	TCAGC	GGTCAGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 19	R	ACTAC	GGACTACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 20	R	CGGAC	GGCGGACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 21	R	GTATA	GGGTATAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 22	R	CTCAT	GGCTCATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 23	R	TTCCA	GGTTCCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P1.2 - 24	R	GGTAA	GGGGTAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Anschließend wurden beide Adapter-Lösungen bei 97,5 °C für 2.5 min im Thermocycler inkubiert und mit einer Rate von 3 °C/min auf 21 °C abgekühlt.

Der P1 Adapter wurde nach der Inkubation auf 1 µM in 50 µL mit ddH_2O verdünnt.

Ligation

Im nächsten Schritt erfolgte die Ligation der verdauten DNA-Proben mit den Adapter P1 und P2. Da der P1 Adapter den Barcode enthält, wurde jeden der 24 Proben ein anderen P1.x Adapter hinzugegeben. Die Zuordnung der P1-Adapter ist in der Tabelle 9 aufgeführt. Der einzelne Adapter P2 wurde bei jeder Probe hinzugefügt.

Tabelle 9: Zuordnung der P1-Adapter zu den DNA-Proben.
Der P1.x-Adapter ordnet den jeweiligen Proben den unterscheidbaren Barcode zu.

Proben ID	Adapter P1
ML023	P1.01
ML025	P1.02
ML029	P1.03
ML030	P1.04
ML032	P1.05
ML033	P1.06
ML034	P1.07
ML035	P1.08
ML050	P1.09
ML051	P1.10
ML061	P1.11
ML073	P1.12
ML088	P1.13
ML089	P1.14
ML090	P1.15
ML091	P1.16
ML098	P1.17
ML102	P1.18
ML104	P1.19
ML107	P1.20
ML110	P1.21
ML111	P1.22
ML112	P1.23
ML114	P1.24

Es wurde mit dem P2-Adapter ein Master Mix hergestellt. Dafür wurde für eine Probe in 19 µL, 8 µL ddH_2O, 2 µL P2 Adapter, 5 µL T4 Puffer 10x und 4 µL *Ligase T4* pipettiert.

In den gesamten Reaktionsansatz für eine Probe wurden 18 µL des Master Mix, 2 µL der vorgesehen P1-Adapter für eine Probe und 30 µL der verdauten DNA-Probe zusammengefügt.

Die Inkubation der Gesamtlösung erfolgte bei 22 °C für 3 Stunden im Thermocycler. Danach wurde die *Ligase T4* für 10 min bei 65 °C inaktiviert.

Da die Proben jetzt durch den Barcode unterscheidbar sind, wurden die Proben alle in einen Eppendorfgefäss gepoolt und aufgereinigt.

Die Aufreinigung und Volumenreduktion wurde bei der ersten Bibliothek mit dem GeneJet PCR Purification Kit durchgeführt. Die zweite Bibliothek wurde mit den magnetischen Beads MagSi-NGS Plus aufgereinigt und das Volumen reduziert. Danach wurde eine erste Größenselektion durchgeführt.

Größenselektion

Die erste Größenselektion wurde mit den magnetischen Beads MagSI-NGS-Plus und der Anweisung für die doppelseitige Selektion durchgeführt. Das Maximum der Fragmentlänge soll bei 300 bp liegen. Als erste wurden die langen Fragmente entfernt, indem 75 % des Probenvolumens (50 µL) an Beads der Probe hinzugefügt wurde. Die kurzen Fragmente wurden mit 65 % des Volumens der Probe an Beads entfernt. Anschließend wurde mit einem Volumen von 38 µL eluiert.

Nach der Ligation der Proben und Aufreinigung und Selektion wurden die Proben wieder mit der Tapestation kontrolliert.

Polymerasekettenreaktion (PCR)

Durch die PCR wird ein Index hinzugefügt, der eine weitere Unterscheidung der Proben ermöglich. Die eingesetzten Primer sind in der Tabelle 10 mit ihren Sequenzen aufgeführt. Der ILLPCR1-Primer wurde für beide Bibliotheken eingesetzt. Die Primer ILLPCR2_01 und ILLPCR2_02 besitzen unterschiedliche Indexe und ermöglichen eine Unterscheidung der zwei DNA-Bibliotheken, wenn die Daten gleichzeitig analysiert werden sollen. Der Primer ILLPCR2_01 wurde für die erste Bibliothek ML1 und der Primer ILLPCR2_02 für die zweiter Bibliothek ML2 eingesetzt.

Tabelle 10: Basensequenzen der Primer.
In der Tabelle sind die Primer-Sequenzen der eingesetzten Primer ILLPCR1, ILLPCR2_1 für ML1 und ILLPCR2_2 für ML2 dargestellt. In Rot sind die sechs Basen, die den Index der Primer bilden gekennzeichnet.

Name	**Primer-Sequnezen**
ILLPCR1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT
ILLPCR2_01	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGC
ILLPCR2_02	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGC

Für die PCR wurde ein Master Mix zusammen pipettiert. Die Zusammensetzung des Master Mix für eine DNA-Probe ist in der Tabelle 11 dargestellt. Das Gesamtvolumen des Reaktionsansatzes beträgt 25 µL. Für die PCR wurde der DNA-Pool separiert, indem 4 µL für jede PCR-Probe eingesetzt wurde. Die Bibliotheken wurde in der Aufreinigung mit 35 µL eluiert. Also wurde die PCR mit 9 Proben und einer negativen Probe mit H_2O durchgeführt. Da Pipettierfehler ausgeglichen werden müssen, wurden 2 Proben extra gezählt. Der Master Mix wurde deshalb für 12 Proben berechnet. [Tab. 11]

Tabelle 11: Master-Mix-Ansatz für die PCR.
Bei 1x sind die Volumina für die Bestanteile des PCR-Ansatzes für eine Probe dargestellt. Bei 10 Proben wurden die Werte bei einer Probe (1x) mit 12 multipliziert.

	1x	**12x**
H_20	13,25 µL	159 µL
Primer 1 ILLPCR 1	1 µL	12 µL
Primer 2 ILLPCR2_01	1 µL	12 µL
PHUSION Buffer HF 5x	5 µL	60 µL
dNTPs	0,5 µL	6 µL
PHUSION Polymerase	0,25 µL	3 µL
DNA-Probe	4 µL	

Der PCR-Lauf im Thermocycler wurde mit folgenden Parametern durchgeführt: Vorheizen mit 98 °C für 30 s. Die Denaturierung für 10 s bei 98 °C, die Anlagerungs-Phase der Primer bei 65 °C für 30 s, die Elongation bei 72 °C für 30 s, das für 20 Zyklen. Danach für 5 min bei 72 °C eine Abschlusselongation und anschließend eine Lagertemperatur bei 4° C.

Größenselektion

Nach der PCR wurden die einzelnen PCR-Produkte wieder zusammengeführt und aufgereinigt. Die erste Bibliothek mit GeneJet PCR Purification Kit und die zweite mit den magnetischen Beads MagSi-NGS Plus. Das aufgereinigte PCR-Produkt wurde mit der Tapestation kontrolliert.

Danach wurde die ddRAD-Bibliotheken mit dem Fragmentierer BluePippin auf eine Größe von 300 bp eingestellt. Die Fragmentierung wurde mit der Anleitung des Geräts BluePippin durchgeführt.

Nach der Fragmentierung wurde die Probe mit der Tapestation kontrolliert und danach zum NGS-Sequenzieren verschickt.

NGS-Sequenzierung der ddRAD-Bibliothek

Die Sequenzierung der ddRAD-Bibliotheken wurde vom Leibniz-Institut für Virologie in der Gruppe Hochdurchsatz-Sequenzierung von Frau Dr. Nakel durchgeführt. Die Sequenzierung erfolgte dort durch ein Illumina MiSeq System. Die NGS-Sequenzierung wurde zweiseitig durchgeführt. Ein Lauf R1 lief vorwärts ab und der zweite Lauf R2 rückwärts.

Bioinformatische Bearbeitung der ddRAD-Bibliothek

Für die bioinformatische Bearbeitung der ddRAD-Bbliothek wurde die Programme FASTQC/MULTIQC, CUTADAPT und STACKS eingesetzt. Für die Populationsanalyse wurde das Programm STRUCTURE und STRUCTURE HARVERSTER genutzt. Die bioinformatischen Bearbeitungen wurden auf den Großrechner vom LIB-Cluster durchgeführt. In der Abbildung 4 ist die Reihenfolge der eingesetzten Programme in ein Grundfließbild dargestellt.

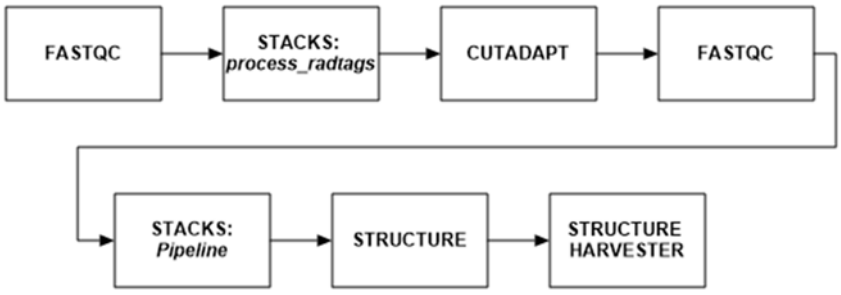

Abbildung 4 Grundfließbild der bioinformatischen Bearbeitung der ddRAD-Binliothek.
Am Anfang werden die die Rohsequenzen mit FASTQC kontrolliert. Danach werden die Sequenzen durch das STACKS-Programm *process_radtags* durch den Barcode zu den Proben geordnet. Als nächste werden die ersten Basen von CUTADAPT aus den Sequenzen geschnitten. Anschließend werden die geschnittenen Sequenzen nochmal mit FASTQC kontrolliert. Jetzt werden die Sequenzen durch die STACKS-Pipline bearbeitet und populationsgenetische Daten erzeugt. Danach werde die Daten eingesetzt, um mit STRUCTURE eine Populationsstruktur zu simulieren. Das Analyse-Programm STRCUTRURE-HARVESTER wird eingesetzt, um ein ΔK zu bestimmen und die finalen Simulationsdaten zu erhalten.

Als erstes wurde das Programm FASTQC ausgeführt damit die Rohsequenzen auf ihre Qualität kontrolliert werden. Für eine bessere Darstellung der Daten wurde das Programm MULTIQC eingesetzt.

Das Programmpaket STACKS wurde eingesetzt, um aus den Rohsequenzen einen Datensatz zusammenzustellen, der für eine populationsgenetische Untersuchung eingesetzt wurde. Die Befehlszeile der STACKS-Programme richten sich nach einem Protokoll von Dr. Oliver Hawlitschek und den STACKS Manual. Die ddRAD-Bibliotheken wurden mit der Standardeinstellung von den einzelnen STACKS - Programmen durchgeführt. Änderung von Standard-Parameter wurde in den Befehlszeilen angegeben. Bei den angebenden Befehlszeilen handelt es sich nur um einen Lauf mit der ersten Bibliothek ML1 aber mit dem Schema wurden auch die zweite Bibliothek und die Zusammenführung der beiden durchgeführt.

Als erstes wurde das Programm *process_radtags* eingesetzt. Das Programm sortiert die Rohdaten nach den eingesetzten Barcodes, reinigt und schneidet die Sequenzen.

Für den Barcode wurde vorher eine Textdatei geschrieben, die in der ersten Spalte die Barcodes enthält und in dem zweiten Spalten, die dazu gehörige Probenbezeichnung. In der Textdatei wurden die Spalten mit der Tab-Taste getrennt. In der Tabelle 12 sind die 24 Barcodes und die Zuordnung zu den 24 Proben aufgeführt.

Tabelle 12: Herstellung der Barcode-Textdatei:
Die linke Spalte ist die Tabelle der Barcodes von der ersten ddRAD-Bibliothek ML1 und rechts ist die Barcode Zuteilung der zweiten Bibliothek ML2 In der Textdatei sind in der ersten Spalte der Barcode und in der zweiten Spalte die Bezeichnung der Probe. Die Trennung der Spalten erfolgt mit der Tab-Taste.

Barcode ML1	Barcode ML2
GCATG ML023	GCATG ML020
AACCA ML025	AACCA ML021
CGATC ML029	CGATC ML026
TCGAT ML030	TCGAT ML027
TGCAT ML032	TGCAT ML036
CAACC ML033	CAACC ML054
GGTTG ML034	GGTTG ML052
AAGGA ML035	AAGGA ML058
ACACA ML050	ACACA ML062
ATACG ML051	ATACG ML070
CTTGG ML061	CTTGG ML071
GAGTC ML073	GAGTC ML075
AGCTA ML088	AGCTA ML076
ACTTC ML089	ACTTA ML077
CATAT ML090	CATAT ML078
CGGCT ML091	CGGCT ML092
CTGAT ML098	CTGAT ML093
GCTGA ML102	GCTGA ML094
GTAGT ML104	GTAGT ML095
GTCCG ML107	GTCCG ML099
TATAC ML110	TATAC ML100
ATGAG ML111	ATGAG ML101
TGGAA ML112	TGGAA ML109
TTACC ML114	TTACC ML115

Das Programm *process_radtags* wurde mit der folgenden Befehlszeile ausgeführt:

process_radtags -P -p RawData -b Barcode_2.txt -o demultiplex_ML1 -t 75 -y fastq /
-r -c -q --disable-rad-check --renz-1 sbfI --renz-2 mseI /
--adapter-1 AGATCGGAAGAGCACACGTCTGAACTCCAGTCA /
--adapter-2 AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT --adapter-mm 2

Das *-P* zeigt an, dass die Rohdaten paired-end-Dateien sind. Das kleine *-p* zeigt den Speicherort der Rohdaten an, *-b* den Speicherort der Barcode-Datei, *-o* zeigt den Speicherort der ausgegebenen Daten an. Der Parameter *-y* gibt das Speicherformat der Ausgabe-Datei an. *-t* gibt die Länge an auf die alle Sequenzen gekürzt werden sollen. *-r* beinhaltet die Rettung von Barcodes und Bindungsstellen von Enzymen. *-- disable-rad-check* verhindert die Suche von enzymatisches Bindungsstellen in den Sequenzen.

Das Programm *process_radtags* erzeugt im Ausgabeordner für jede Probe vier Dateien. Zwei bearbeitete Dateien für die Rohdaten .R1 und .R2 mit den Endungen .1.fq und .2.fq und zwei Datensätzen die entfernten Sequenzen enthalten mit rem.1.fq und rem.2.fq.

Um die Sequenzen die 10 ersten Basenpaare zu entfernt, die durch die Sequenzierung eine starke Variabilität der Basen aufweist, wurde das Programm CUTADAPT eingesetzt. Die Befehlszeile lautet:

```
Name="ML023 ML025 ML029 ML030 ML032 ML033 ML034 ML035 ML050 ML051
ML061 ML073 ML088 ML089 ML090 ML091 ML098 ML102 ML104 ML107 ML110
ML111 ML112 ML114"

for Probe in $Name
do
        cutadapt -u 10 -U 10 -o demultiplex_ML1_cut/${Probe}.1.fq /
        -p demultiplex_ML1_cut/${Probe}.2.fq demultiplex_ML1/${Probe}.1.fq /
        demultiplex_ML1/${Probe}.2.fq /
done
```

Der Parameter *-u* schneidet die ersten Basenpaare der .1.fq-Dateien ab und *-U* die ersten 10 bp der paired-end-Datei ab.

Nach der Bearbeitung der Sequenzen mit CUTADAPT wurden die Sequenzen mit FASTQC kontrolliert.

Anschließend wurden die Proben entfernt, die über oder unter präsentiert waren. Um diese Proben zu ermitteln, wurde die Anzahl der Sequenzen pro Probe gemittelt und die

Standardabweichung vom Mittelwert berechnet. Durch das Hinzuzählen und Abziehen der Standardabweichung vom Mittelwert wurde ein Bereich bestimmt, indem die Proben liegen, sollten die weiterverarbeitet wurden. In den Tabellen 17 und 18 sind die Werte dargestellt. Bei der ersten Bibliothek wurden 7 Proben (ML025, ML029, ML033, ML050, ML051, ML091, ML111) entfernt, bei der zweiten ML2 wurden 2 (ML036, ML078) entfernt und bei der Zusammenführung Ml1 + ML2 wurden 11 Proben (ML025, ML036, ML050, ML051, ML061, ML078, ML091, ML095, ML098, ML107, ML111) entfernt.

Bei dem nächsten Schritt wurde die STACKS-Pipeline genutzt. Aus den bearbeiteten Daten wurde mit dem Programm *ustacks* die Loci gebildet. Für das Programm *ustacks* wurden nur die single-end-Daten mit der Endung .1.fq genutzt. Für das Programm *ustacks* wurde ein Skript mit eine *for*-Schleife geschrieben. Die Schleife durchläuft alle Dateien im Ordner *demultiplex_ML1* mit den Suffix .1.fq und speichert die Ausgabe in Ordner *process_samples_ML1* ab.

```
Name="ML023 ML025 ML029 ML030 ML032 ML033 ML034 ML035 ML050 ML051
ML061 ML073 ML088 ML089 ML090 ML091 ML098 ML102 ML104 ML107 ML110
ML111 ML112 ML114"

x=1
for Probe in $Name
do
        ustacks -f demultiplex_ML1_cut/${Probe}.1.fq -i $x -o process_samples_ML1
        --name $Probe
        let "x+=1"
done
```

Mit dem Parameter *--name* wurde die Ausgabedatei nur mit dem Namen der Probe gespeichert. *-i* bezeichnet die ID der Proben.

Mit den Programm *cstacks* wurden die Loci katalogisiert und zusammengefasst.

```
cstacks -P process_samples_ML1 -M Popmap.txt -n 2
```

Der Parameter -n gibt an wieviel Unterschiede zwischen den Loci der Proben erlaubt sind, um diese zusammenzufassen. Für weiteren Schritte wurde eine Populationsmappe benötigt. Für die Mappe wurde eine Text-Datei mit der Zuteilung der Proben zu einer Population geschrieben. In der ersten Spalte wurde die Proben ID geschrieben und in die zweite Spalte

der Herkunftsort der Probe. In der Tabelle 13 sind die Zuordnungen dargestellt. Der Parameter -P bezeichnet den Speicherordner der genutzten Daten und gleich der Speicherordner der erzeugten Daten. -M gibt den Speicherort der Populationsmappe an.

Tabelle 13: Populationsmappe von der beiden ddRAD-Bibliothek ML1 und ML2.
In der Tabelle sind die Proben dargestellt, die in die jeweiligen Populationsmappen geschrieben wurden bzw. für eine weiter Analyse mit der STACKS-Pipeline eingesetzt wurden. Die ganze Zahl, die durch einen Tab von der Proben-ID getrennt ist, stellt den Fundort als Population da. (1) ist SVK, (2) ist HUN, (3) ist DEUK, (4) ist DEUR, (5) ist NLD, (6) ist kA und (7) ist DEUM

Populationsmappe ML1	Populationsmappe ML2
ML023 1	ML020 1
ML030 1	ML021 1
ML032 2	ML026 1
ML034 2	ML027 1
ML035 2	ML054 3
ML061 3	ML052 3
ML073 4	ML058 3
ML088 5	ML062 3
ML089 5	ML070 4
ML090 5	ML071 4
ML098 5	ML075 4
ML102 6	ML076 4
ML104 7	ML077 4
ML107 7	ML092 5
ML110 4	ML093 5
ML112 4	ML094 5
ML114 4	ML095 5
	ML099 6
	ML100 6
	ML101 6
	ML109 5
	ML115 5

Mit dem Programm *sstacks* wurden die zusammengefassten Loci mit den Loci der einzelnen Proben gepaart, um die Haplotypen der Loci zu bestimmen.

sstacks -P process_samples_ML1 -M Popmap.txt

Das Programm *tsv2bam* konventiert die tsv-Dateien die von *sstacks* erzeugt wurden in *bam*-Dateien um.

tsv2bam -P process_samples_ML1 -M Popmap.txt -R demultiplex_ML1_cut

Das *gstacks* Programm genotypisiert die Daten und fügt die paired end-Reads hinzu.

gstacks -P process_samples_ML1 -M Popmap.txt --rm-pcr-duplicates

Bei den Parameter -P, -M, -R sind die Angaben für die Speicherort der benötigten Daten. Bei *gstacks* wurde noch der Befehl *--rm-pcr-dublicates* hinzugefügt, um die PCR-Dublicate bzw. um Readpaare mit der gleichen Insert-Länge zu entfernen

Das Populationsprogramm *populations* fasst alle Ausgaben der anderen Programme zusammen und analysiert die Population der eingesetzten Proben. Das Programm bildet genetische Statistiken und speichert das in bestimmt Speicherformaten ab. Desweitern filtert das Programm die Daten nach der Häufigkeit der Loci die in einer Population vorkommen sollen.

Populations -P process_samples_ML1 -M Popmap.txt -O results_populations_ML1

--write-single-snp --fstats --fasta-samples --plink --structure -r 0.5 -p 1

Die Parameter *-P* und *-M* sind die Speicherordner bzw. Datei der genutzten Daten. *-O* gibt den Speicherordner an in dem die berechneten Daten gespeichert wurden. Bei der Einstellung *--write-single-snp* wurde nur die erste SNP pro Loci analysiert. *--fstats* aktiviert die F-Statistik für die SNP und Haplotypen. Bei *--fasta-samples* wurden die Ergebnisse in dem Fasta-Format gespeichert. *--plink* gibt die Ausgabe in ein Plink-Format wieder. *--structure* speichert die Ausgabe in structure-Format ab. *-r* gibt den Prozentsatz an, der benötigt wurde, um einen Locus in die Population der Probe einzuarbeiten. *-p* gibt die minimale Anzahl eines Locus in einer Population an, um diesen Loci einzufügen.

Populationsanalyse mit STRUCTURE

Die Ausgaben von den Programm *populations* von STACKS wurde eingesetzt, um eine Populationsstruktur zu erstellen. Dafür wurden verschieden Population (K) berechnet, um eine stabile Population ΔK zu bestimmen. Für die Bestimmung von ΔK wurden für die beiden Bibliotheken 7 Population K berechnet. Für jedes K wurden 3 Läufe berechnet. Für diese Berechnungen wurde eine Anlaufzeit (burnin) von 5000 und eine MCMC-Wiederholung von 10000 eingestellt. Anschließenden wurde die Ausgabe mit dem Internet-Programm STRUCTURE HARVESTER analysiert und das ΔK bestimmt. Das ΔK wurde mit der Evanno-Methode bestimmt. Mit den bestimmten ΔK wurde eine neue Berechnung gestartet mit einer längeren Laufzeit. Die Anlaufzeit betrug 50000 und die MCMC-Wiederholungen lagen bei 100000.

Ergebnisse

Messergebnisse der ddRAD-Bibliothek

In der Tabelle 14 sind die Konzentrationsmessung der ddRAD-Bibliothek ML1 aufgeführt. Die ersten Spalte ist die Proben-ID und die anderen drei Spalten enthalten die Konzentrationsmessung der einzelnen Proben nach jeden Bearbeitungsschritt. Als erstes die Extraktion, dann der Verdau und als letztes die Ligation. Der Wert für die PCR ist nur einer, weil nach der Ligation die Proben zusammengeführt wurden. Die letzte Konzentrationsangabe ist die Messung der Tapestation bei der nur der Peaks ausgemessen wurde.

Tabelle 14: Konzentration der Ersten Bibiliothek ML1.
Es sind die DNA-Konzentrationen der ML1-Bibliothek nach der Extraktion, nach dem Verdau und nach der Ligation dargestellt. Die DNA-Konzentration wurde mit den Nanodrop bestimmt.

Proben-ID	Konz. Extraktion ng/µl	Konz. Verdau ng/µL	Konz. Ligation ng/µl
ML023	77	19,4	26,1
ML025	93,7	19,1	27
ML029	88,7	18,1	24
ML030	96,7	16,7	26,2
ML032	50,2	18,1	26,8
ML033	84	17,6	24
ML034	122,2	20,4	27,4
ML035	102,1	17,1	26,6
ML050	158,6	15	23,6
ML051	175,8	23,2	23,3
ML061	193,3	36,2	23,9
ML073	303,8	20,5	22,6
ML088	281,4	19,1	29
ML089	572,5	15,1	23,3
ML090	211,5	11,2	23,3
ML091	214,5	18,8	25,6
ML098	329	21,2	26,1
ML102	182,2	13,6	26,1
ML104	104,2	21	31,1
ML107	129,5	15,8	20,9
ML110	109,7	12,8	25,9
ML111	110,4	12	21,1
ML112	106,8	13,9	28,4
ML114	60,8	14	24

Bei der Konzentration der Extraktion liegen die die Proben zwischen 60,8 ng/µL bei ML114 und 572,5 ng/µL bei ML089. Die meisten Proben liegen um 100 ng/µL. Nach der Digestion liegen die meisten Proben zwischen 12,8 ng/µL und 23,2 ng/µL. Nur bei der Probe ML61 liegt

die Konzentration mit 36,2 ng/µL höher. Der Konzentration der Proben liegt nach der Ligation zwischen 21,1 ng/µL und 31,1 ng/µL. Die Konzentration des PCR-Pools liegt bei 167 ng/µL.

In der Tabelle 15 sind die Konzentrationswerte von den Proben der Bibliothek ML2. Die Tabelle zeigt nur die Extraktionswerte an. Die einzelnen Schritte wurden bei ML2 nicht gemessen, außer die Konzentration nach der PCR mit 138,1 ng/µL und nach der Größenselektion mit einer Konzentration von 17 ng/µL. Die Konzentration der Extraktion liege zwischen 71,3 ng/µL und 322,3 ng/µL verteilt. Diese Werte wurden für die Einstellung der DNA-Menge für Herstellung der ddRAD-Bibliothek genommen.

Tabelle 15: DNA-Konzentrations der zweiten Bibliothek ML2.
Für die zweite Bibliothek ML2 sind die DNA-Konzentration der Extraktion dargestellt. Nach diesen Werten wurde die Menge von 100 ng eingestellt. DNA-Konzentration wurde mit den Nanodrop bestimmt.

Proben-ID	Konz. Extraktion ng/µl
ML020	72
ML021	96,6
ML026	95,1
ML027	235
ML036	120,1
ML052	112,1
ML054	118,7
ML058	144,2
ML062	140,2
ML070	227,4
ML071	268,6
ML075	159,5
ML076	183,6
ML077	138,9
ML078	165,2
ML092	313,8
ML093	322,3
ML094	263,5
ML095	178,7
ML099	298,7
ML100	270,1
ML101	71,3
ML109	80,1
ML115	82,8

Kontrolle der ddRAD-Bibliothek durch die Tapestation

In den Abbildung 5 bis 8 sind die Gel-Bilder der Tapestation von der ersten ddRAD-Bibliothek ML1, durchgeführt für jeden Schritt der bei der Herstellung wurden, dargestellt.

In der Abbildung 5 sind die Ursprungsproben aufgezeichnet. Es sind schwache Banden bei allen Proben unterhalb von 2500 bp zuerkennen. Zwischen 3500 bp - 5000 bp sind bei den Proben Banden zu sehen. Da die DNA noch hochmolekular ist, wird die Verteilung der DNA von den eingesetzten Screentape D5000 nicht richtig dargestellt, weil die DNA die größer als 5000 bp ist, nicht aufgezeichnet wird. Die Banden bei der grünen und lila Linie sind Lower und Upper Banden, die durch den Dyer, der den Proben zugegeben wurden, stammen.

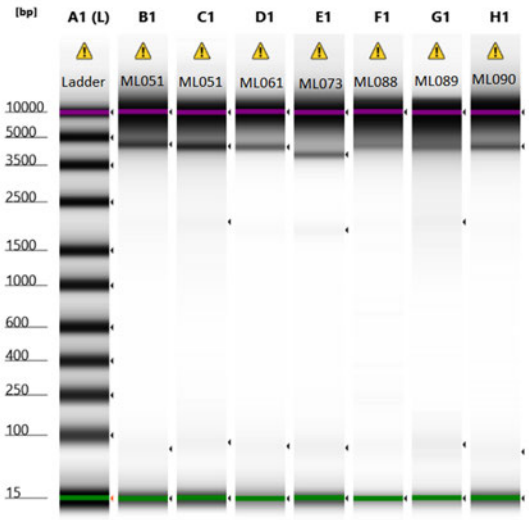

Abbildung 5: Tapestation-Aufnahme der Ursprungsproben von ML1.
Die Abbildung zeigt das Tapestation-Laufbilder von den Proben ML50, ML51, ML61, ML73, ML88, ML89 und ML90. Der Lauf wurde mit dem Screentape D5000 durchgeführt. Die kleinen Pfeile an der Seite der Spalten zeigt einen Peak an. Der Ladder hat eine Range von 15 bp bis 10000 bp. Das gelbe Zeichen ist ein Warnzeichen und zeigt an, dass das Screentape das Ablaufdatum überschritten hat.

Die Abbildung 6 stellt die Proben nach der Digestion mit *SbfI* und *MseI* dar. Die Proben wurde mit den Screentape D5000 gemessen. Die Abbildung wurde im Analyse-Programm der Tapestation vergrößert (Scale-down), um die Banden sichtbarer zu machen. Bei den Proben ML51, ML73, ML89 sind DNA-Fragmente von der Größe von 1000 bis 100 bp zu sehen. Der größte Peak wird durch einen kleinen Pfeil an der Seite der Laufbahn angezeigt. Der Peak liegt bei den genannten Proben um die 250 bp. Die Laufbahn der Probe ML91 wurde nicht richtig detektiert von der Tapestation. Das gelbe Warnzeichen zeigt an, dass das benutzte Screentape das Ablaufdatum überschritten hat.

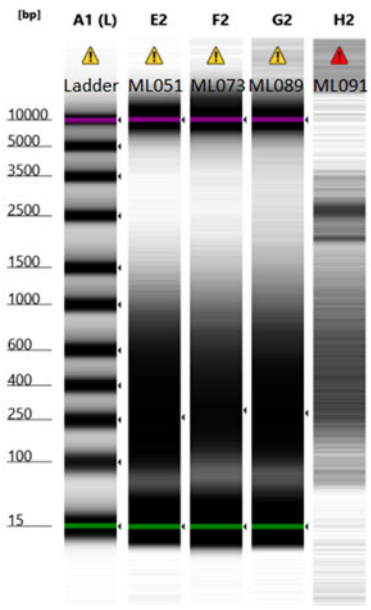

Abbildung 6: Tapestation - Aufnahme der Digestion von ML1.
Es sind die Proben ML51, ML73, ML89 und ML91 dargestellt. Für die Aufnahmen wurde das Screentape D5000 eingesetzt. Das rote Warnzeichen zeigt an, dass der Lauf der Probe ML091 von der Tapestation nicht richtig erkannt wurde.

In der Abbildung 7 sind die einzelnen Proben nach der Ligation zu sehen. Die Proben wurde mit dem Screentape D1000 gemessen. Dieses Tape stellt nur die Banden unterhalb von 1500 Bp da. Unterhalb von 100 Bp sind zwei Peaks zuerkennen außer bei der Probe ML33. Die Peaks kommen von nicht angelagertem Adapter. Die größte Verteilung von DNA-Fragmente ist bei allen Proben, um die 400 Bp zu erkenne.

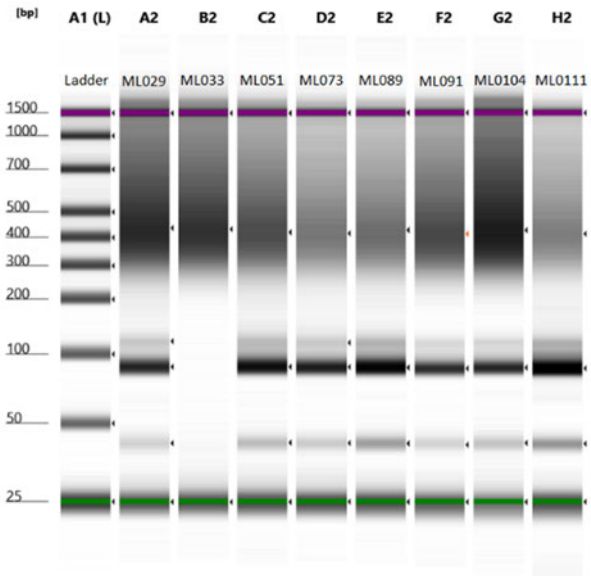

Abbildung 7: Tapestation - Aufnahme der Ligation von ML1.
Für diese Aufnahme wurde das Screentape HS D1000 benutzt. Dabei wurden die folgenden Proben von ML1 gemessen: ML29, ML33, ML51, ML73, ML89, ML91, ML104 und ML111.

Die Tapestation-Aufnahme vom PCR-Pool und nach der Größenselektion sind in der Abbildung 8 dargestellt. Der PCR-Pool unterscheidet sich nicht von der Darstellung der Proben nach der Ligation. Da beim PCR alle Proben zusammengefügt wurden, ist die DNA-Verteilung deutlicher dargestellt als bei den einzelnen Proben. Nach der Größenselektion mit dem BluePippin hat die Gellaufbahn nur einen Peak bei 300 Bp.

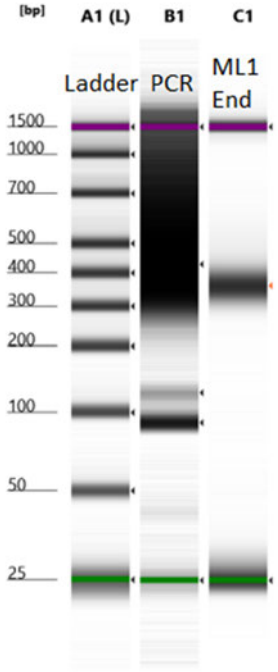

Abbildung 8: Tapestation - Aufnahme vom PCR-Lauf und der Größenselektion von ML1.
Für die Aufnahme wurde das Screentape HS D1000 eingesetzt. Die Aufnahme wurde mit der PCR-Pool und den selektierte Probe ML1 End durchgeführt.

Die Kontrolle der zweiten Bibliothek ML2 durch die Tapestation ist in den Abbildungen 9 bis 12 dargestellt. Bei den Abbildungen handelt es sich um die Grafische Darstellung der Tapestation. Die horizontale Achse zeigt die Länge der DNA in Basenpaaren (bp) an die vertikale Achse zeigt die Intensität der Probe an.

In der Abbildung 9 wird die Digestion der Probe ML054 die DNA-Verteilung hat bei 293 Bp den größten Peak und bei 50 bp einen kleineren Peak. Der Peak, der bei 1987 bp liegt, ist außerhalb der Range vom Screentape D1000 und wird deswegen nicht betrachtet. Die Grafik ist mit dem Analyse-Programm vergrößert worden, um die Darstellung zu verbessern. Die DNA-Konzentration des Peaks von ML1 End, durch die Tapestation bestimmt, liegt bei 569 pg/µL

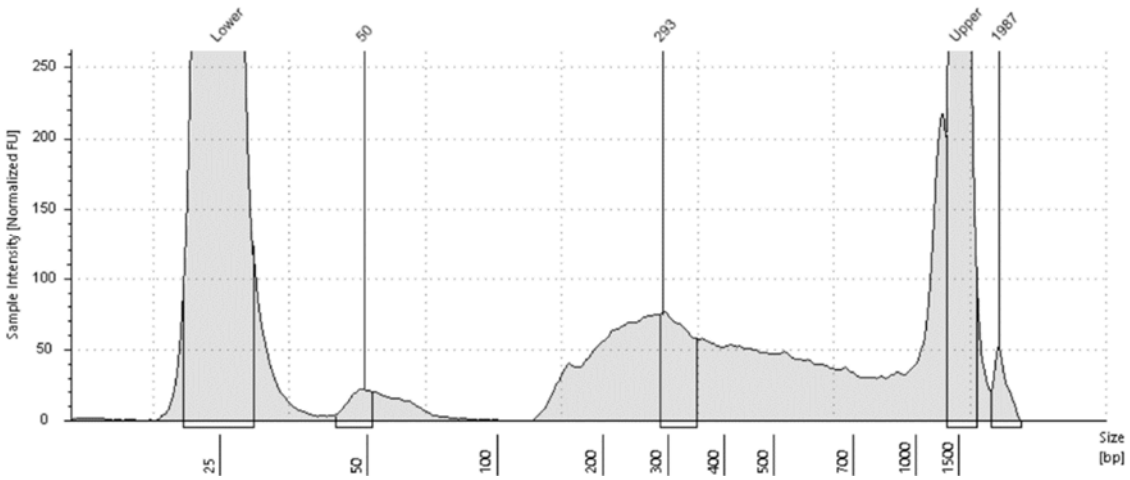

Abbildung 9: Tapestation-Kurve der Digestion der Probe ML054 von ML2.
Für diesen Lauf wurde das Screentape D1000 eingesetzt. Die Kurve wurde herangezoomt, um die Peaks besser zu sehen. Die Lower und Upper Peaks sind vom hinzugefügten Dyer. Die Peaks innerhalb der Range liegen bei 50 bp und 293 bp.

Die Abbildung 10 zeigt den Pool nach der Ligation der Adapter und der Größenselektion mit den magn. Beads an. Der größte Peak liegt in dieser DNA-Verteilung bei 365 bp und einen kleinen Peak bei 123 bp. Zwischen 365 bp und 1500 bp fällt die DNA-Verteilung ab.

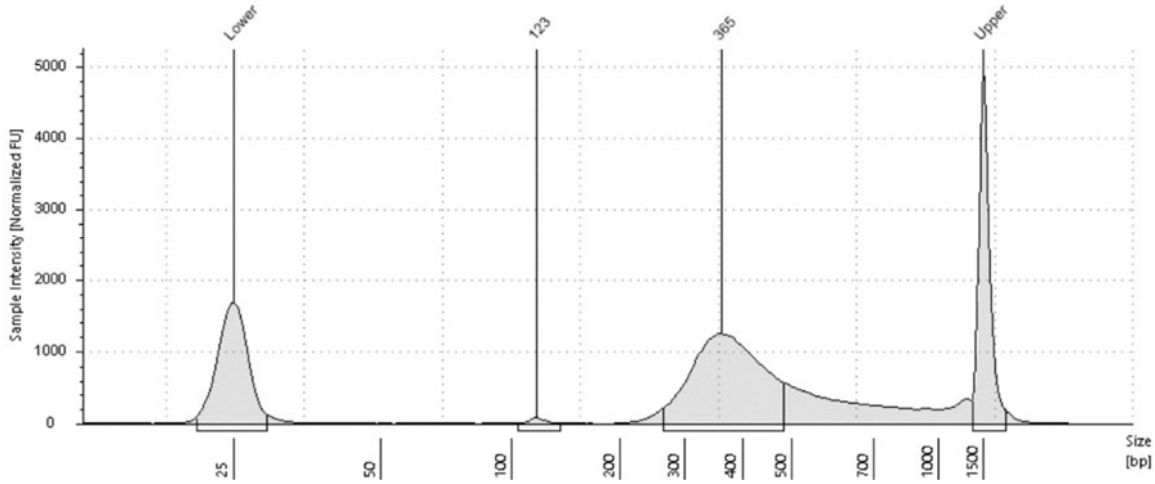

Abbildung 10: Tapestation-Kurve nach der Ligation und Größenselektion vom ML2-Pool.
Für den Lauf wurde das Screentape D1000 benutzt. Die Kurve zeigt einen Peak bei 123 bp und einen Peak bei 365 bp an.

In der Abbildung 11 wird der Tapestation-Lauf mit dem PCR-Pool dargestellt. Ein kleiner Peak liegt bei 135 bp. Zwei größere Peaks liegen bei 291 bp und 339 bp.

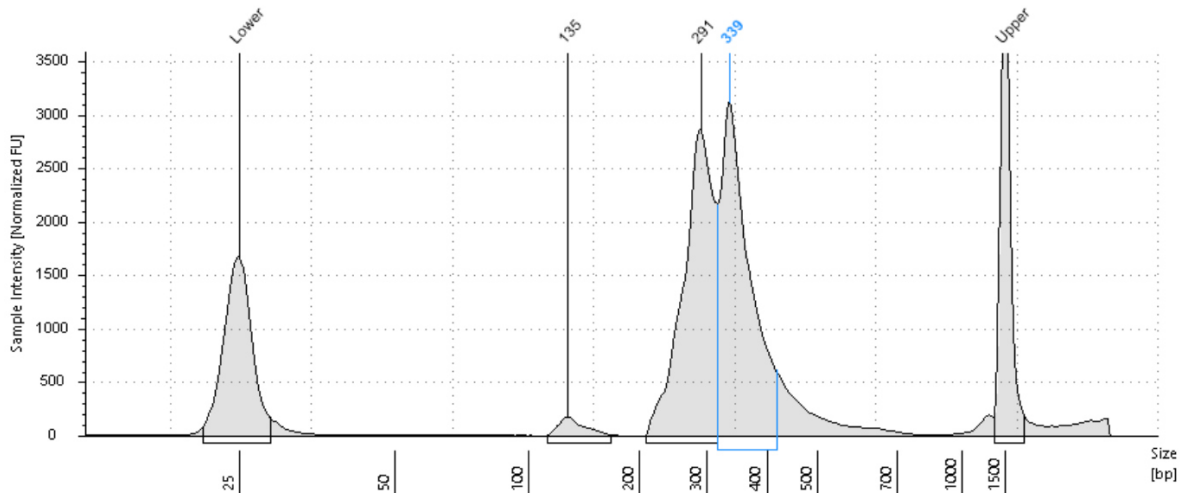

Abbildung 11: Tapestation-Kurve nach der PCR von ML2.
Es wurde bei diesem Lauf das Screentape D1000 eingesetzt. Ein geringer Peak liegt bei der Position 135 bp und ein Doppel-Peak liegt zwischen 291bp und 339 bp.

Die DNA-Verteilung nach der Größenselektion mit BluePippin ist in der Abbildung 12 dargestellt. In dieser Verteilung gibt es nur einen Peak, der bei 298 bp liegt. Die Grafik wurde mit den Analyse Programm vergrößert dargestellt. Die Messung der DNA-Konzentration durch die Tapestation hat für den Peak bei 298 bp eine Konzentration von 3,27 ng/µL

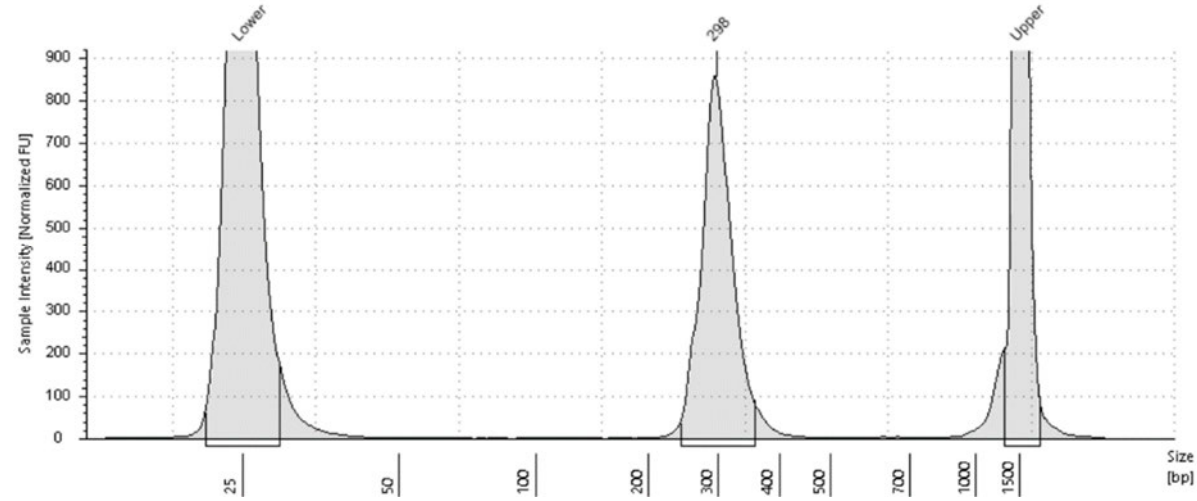

Abbildung 12: Tapestation-Kurve der Größenselektion mit Bluepippen von ML2.
Für den Tapestation-Lauf wurde das Screentape D1000 benutzt. Die Kurve besitzt nur einen gleichförmigen Peak bei 298 bp. Die Kurve wurde herangezoomt, um die Kurve besser darzustellen.

Befund der Qualitätsanalyse der Rohdaten

Die Qualitätsanalyse wurde mit den Programm FASTQC durchgeführt. In der Tabelle 16 sind die Basisdaten der Rohdaten von den beiden Bibliotheken ML1 und ML2 aufgeführt. In den Abbildung 13 und 14 ist die Betrachtung der Sequenzen durch den Phred-Qualitäts-Wert dargestellt. Die Abbildung 15 und 16 zeigen die Basenverteilung über die Basensequenz an. Die Werte in der Tabelle und die Abbildungen stammen von den Programm MULTIQC.

Um die Verbesserung der Qualität zu beurteilen, sind die einzelnen Proben der Bibliotheken nach dem Reinigen und Schneiden mit FASTQC kontrolliert worden und die Ergebnisse in der Tabelle 17 und in den Abbildung 17 bis 20 dargestellt.

In der Tabelle 16 sind die Qualitätsbasisdaten der jeweiligen Sequenzierungsrichtungen der beiden Bibliotheken aufgeführt. R1 bezeichnet die Vorwärtsrichtung und R2 die Rückwärtsrichtung der Sequenzierung. Die Readlänge bei allen vieren geht bis zu 151 bp. Der GC-Gehalt liegt nur bei ML2_R1 bei 48 %, bei den anderen dreien liegt der Gehalt bei 49 %. Bei der Anzahl der Sequenzen unterscheiden sich die beiden ddRAD-Bibliotheken. ML1 hat 172,5 M Sequenzen und ML2 hat 126 M Sequenzen. Bei den Werten zu PCR-Duplikation liegt der Wert für beide Richtung von ML2 bei 93 %. Die Duplikation von ML1 ist um die 20 % geringer als bei ML2, aber die beiden Sequenzierungsrichtungen unterscheiden sich um 10 %. Die Anteile der überrepräsentierten Sequenzen ist bei ML1 geringer mit 0 % (R1) und 3,24 % (R2) als die Anteile von ML2 die bei 14,31 % (R1) und 21,26 % (R2) liegen.

Tabelle 16: Qualitätsdaten der Rohsequenzen von ML1 und ML2.
Die Tabelle zeigt die Basiswert an die von FASTQC und MULTIQC ausgegeben werden. Anzahl der Seqeunzen in M, die Readlänge in Basenpaare (bp), GC-Gehalt in Prozenten %, die PCR-Duplikation und Überrepräsentierte Sequenzen in Prozenten % von der Gesamtanzahl der Sequenzen.

Datenname	Anzahl der Seq.	Readlänge	GC-Gehalt	Duplikation	Überrepräs. Seq.
ML1_R1	172,5 mio.	151 bp	49 %	62,4 %	0 %
ML1_R2	172,5 mio.	151 bp	49 %	73,3 %	3,24 %
ML2_R1	126 mio.	151 bp	49 %	93 %	14,32 %
ML2_R2	126 mio.	151 bp	48 %	93,7 %	21,26 %

In der Abbildung 13 wird der Phred-Qualitäts-Wert über alle Basenpaare aufgetragen. Alle vier Sequenzdaten liegen im grünen Bereich über einen Wert von 28. Der Anfang (10 bis 12 bp) und das Ende der Sequenzen (letzten 10 bp) verzeichnen einen Abfall des Qualitätswertes bis zu 28 oder knapp unter diesen Wert, sonst liegen der Werte von den meisten Basenpaarposition von 30 bis 35.

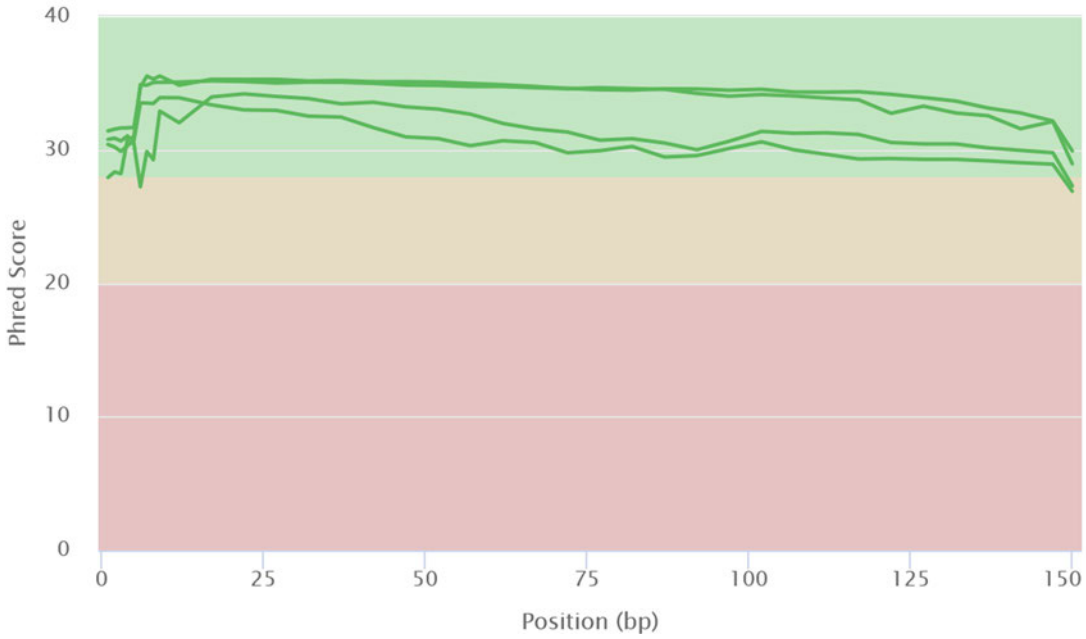

Abbildung 13: Phred-Qualitätswert der Rohdaten von ML1 und ML2.
In dieser Abbildung ist der Phred-Qualität-Wert über alle Basenpaare dargestellt. Die Werte von den allen Reads (R1, R2) liegen im grünen Bereich zwischen 28 und 35. Die Enden fallen etwas unter 28. Die Abbildung wurde vom Programm MULTIQC erstellt.

In der Abbildung 14 wird der Phred-Wert pro Sequenzen dargestellt. Diese Darstellungen zeigt die Anzahl der Sequenzen an, die einen bestimmten Phred-Wert haben. Die Kurven der zweiten Bibliothek geht erst im grünen Bereich ca. 33 exponentiell nach oben bis auf 35. Die maximale Anzahl der Reads an dem Wert von 35 liegt bei 78 mio. Reads bei ML2_R1. Vorher gibt es einen schwachen Anstieg der Sequenzen mit einem schwachen Wert von 20 bis 28. Bei der ersten Bibliothek steigen die Kurven ab einen Wert von 20 kontinuierlich bis auf einen Wert von 34 und fällt danach wieder ab. Das Maximum bei ML1_R1 liegt bei 31 mio. Reads bei einem Phred-Wert von 34.

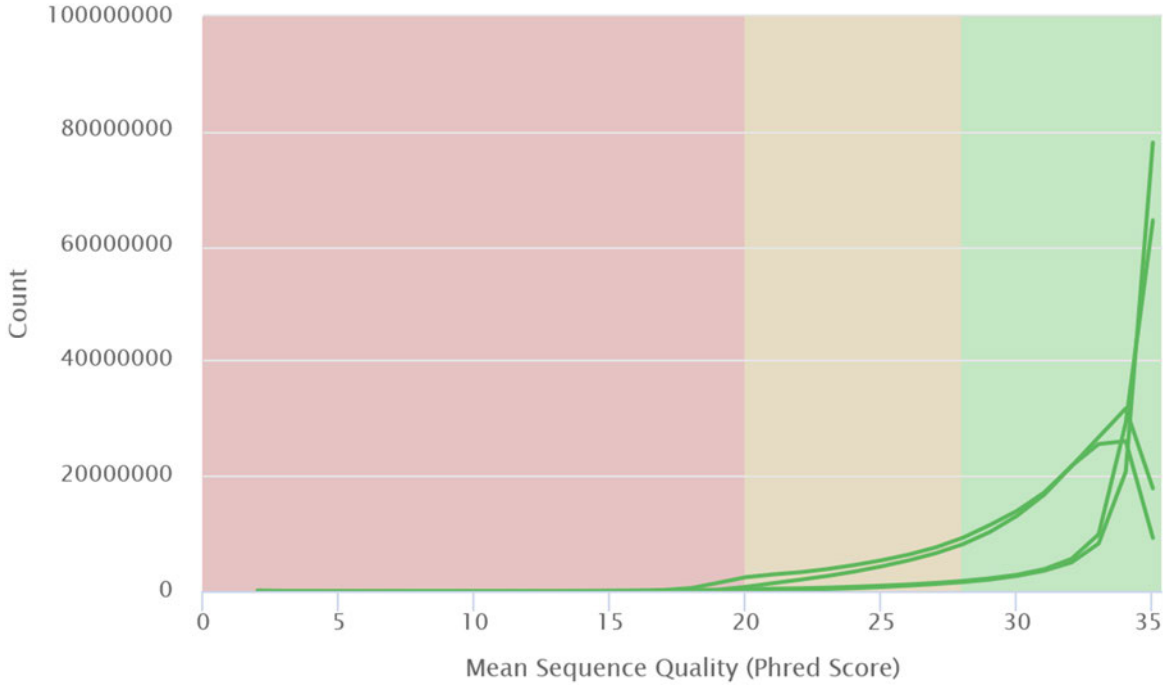

Abbildung 14: Phred-Qualitätswert pro Sequenzen von ML1 und ML2.
Die Darstellung zeigt den Phred-Qualitäts-Wert über alle Sequenzen. Die Sequenzen liegen zum größten Teil bei einem Wert zwischen 28 und 35. Die Abbildung wurde von MULTIQC erstellt.

In den Abbildung 15 und 16 ist die Basenverteilung in den Reads von ML1_R1 und ML2_R1 dargestellt. Bei beiden Darstellungen weist die Basenverteilung eine große Variabilität auf in den ersten 10 bp. In diesen Bereich schwanken die Basenanteil von 50 % bei ML1_R1 und bei ML2_R1 bis zu 80 %. Die Kurven zeigen über die gesamten 150 bp eine starke Variabilität zwischen den jeweiligen Basen an einer Basenpaarposition. Z.B an der Position 60 bp von ML1_R1 liegen die Basen Anteile: T bei 36,7 %, A bei 19,4 %, G bei 18,1 % und C bei 25,8 %. Bei der gleichen Position bei ML2_R1 sind die Basenanteile T 27,6 %, A bei 22,1 %, G bei 33,7 % und C bei 16,5 %.

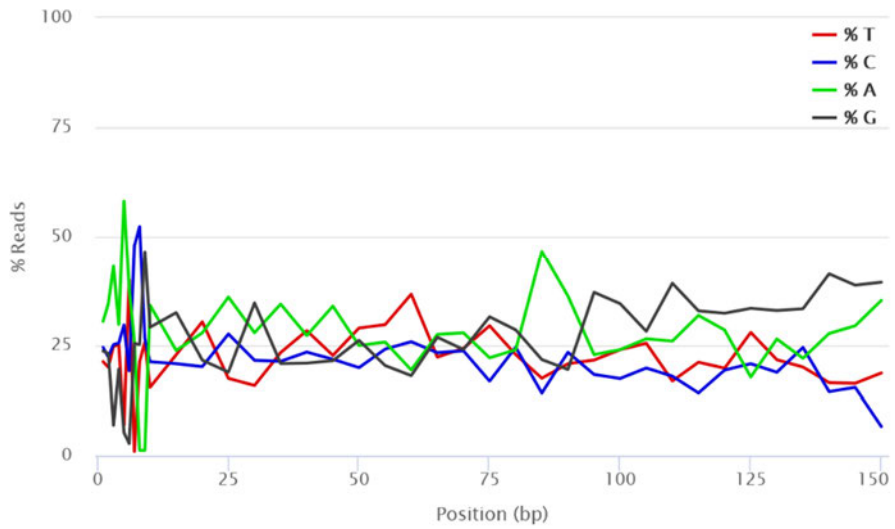

Abbildung 15: Basenverteilung der Reads bei ML1_R1.
Darstellung der Basenverteilung der Reads in Prozenten % in Bezug auf die Basenposition. Die Abbildung wurde vom Programm MULTIQC erstellt.

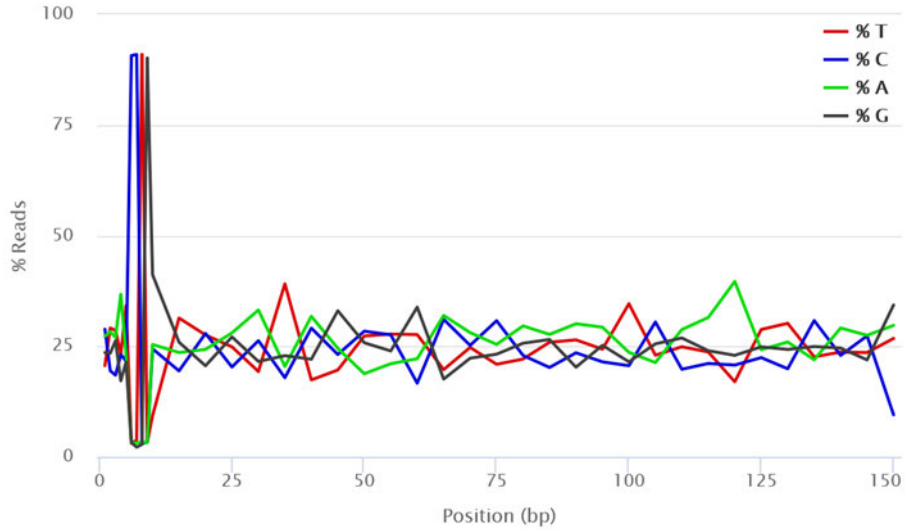

Abbildung 16: Basenverteilung der Reads bei ML2_R1.
Darstellung der Basenverteilung der Reads in Prozenten % in Bezug auf die Basenposition. Die Abbildung wurde vom Programm MULTIQC erstellt.

In der Tabelle 17 sind die Qualitätsdaten von allen Proben aus den Daten von ML1_R1 und ML2_R2 nach der Bearbeitung mit *process_radtags* und CUTADAPT dargestellt. Die Anzahl der Sequenzen reicht von 0,8 Mio. bis zu 13,5 Mio. Die Readlänge liegt bei allen Proben bei 65 Bp nach dem Entfernen der ersten 10 Bp. Der GC-Gehalt liegt zwischen 44 % und 49 %. Die Duplikation unterscheiden sich die Beiden Bibliotheken voneinander die Proben von ML2 liegen um die 90 % und die Proben von ML1 liegt die Duplikation 65 % bis 80 %. Die Anteile der überrepräsentierten Sequenzen liegen bei den einzelnen Proben zwischen 8 % und 21 %. Vorbei die Proben der zweite Bibliothek ML2 den höheren Anteil besitzen.

Tabelle 17: Qualitätsdaten nach den Schneiden mit CUTADAPT.

Die Tabelle zeigt die Basisdaten, die von FASTQC und MULTIQC erzeugt werden. Es werden die Bibliothek und Proben ID angezeigt, die PCR-Duplikation-Anteile, der Basengehalt von GC, die Länge der Reads, die Menge der Sequenzen und die Anteile der überrepräsentierten Sequenzen.

Bibl.-ID	Proben ID	Duplikation	GC-Gehalt	Readlänge	Seq. Anzahl Mio.	Überreprä. Seq.
ML2	ML020	91,60 %	49 %	65 bp	3,7	18,91 %
ML2	ML021	90,10 %	49 %	65 bp	4,2	17,94 %
ML1	ML023	76,60 %	47 %	65 bp	2,6	11,34 %
ML1	ML025	80,60 %	45 %	65 bp	8	15,77 %
ML2	ML026	91,30 %	49 %	65 bp	5	19,98 %
ML2	ML027	91,70 %	49 %	65 bp	4,9	19,69 %
ML1	ML029	81,80 %	46 %	65 bp	5,2	14,38 %
ML1	ML030	75,60 %	48 %	65 bp	2,6	8,39 %
ML1	ML032	75,40 %	47 %	65 bp	2,5	9,22 %
ML1	ML033	81,60 %	45 %	65 bp	5,7	19,68 %
ML1	ML034	75,60 %	47 %	65 bp	2,3	8,88 %
ML1	ML035	76,50 %	46 %	65 bp	3,3	11,13 %
ML2	ML036	92,90 %	49 %	65 bp	6,9	19,75 %
ML1	ML050	80,00 %	44 %	65 bp	6,2	18,68 %
ML1	ML051	69,30 %	47 %	65 bp	1,4	9,29 %
ML2	ML052	91,30 %	49 %	65 bp	2,8	20,59 %
ML2	ML054	88,80 %	49 %	65 bp	3	19,05 %
ML2	ML058	90,60 %	49 %	65 bp	3,5	19,86 %
ML1	ML061	74,70 %	47 %	65 bp	1,8	9,52 %
ML2	ML062	90,40 %	49 %	65 bp	4,5	21,14 %
ML2	ML070	90,70 %	49 %	65 bp	3,7	19,18 %
ML2	ML071	92,50 %	49 %	65 bp	4,4	20,04 %
ML1	ML073	73,80 %	45 %	65 bp	2	14,37 %
ML2	ML075	92,50 %	49 %	65 bp	4,9	21,36 %
ML2	ML076	91,90 %	49 %	65 bp	4,7	20,67 %
ML2	ML077	91,90 %	49 %	65 bp	4,7	20,47 %
ML2	ML078	93,60 %	49 %	65 bp	13,5	21,69 %
ML1	ML088	79,00 %	46 %	65 bp	4,2	13,13 %
ML1	ML089	79,50 %	45 %	65 bp	4	17,19 %
ML1	ML090	72,40 %	46 %	65 bp	2,2	10,68 %
ML1	ML091	70,40 %	48 %	65 bp	1,4	8,65 %
ML2	ML092	90,20 %	49 %	65 bp	3,3	20,53 %
ML2	ML093	92,20 %	49 %	65 bp	4,8	20,37 %
ML2	ML094	91,70 %	49 %	65 bp	4,4	20,71 %
ML2	ML095	92,50 %	49 %	65 bp	6,7	20,74 %
ML1	ML098	74,80 %	47 %	65 bp	1,7	8,94 %
ML2	ML099	92,00 %	49 %	65 bp	4,7	21,66 %
ML2	ML100	89,30 %	49 %	65 bp	3	18 %
ML2	ML101	91,80 %	49 %	65 bp	5,3	20,55 %
ML1	ML102	79,70 %	45 %	65 bp	3,4	15,76 %
ML1	ML104	75,10 %	47 %	65 bp	2,5	9,65 %
ML1	ML107	71,00 %	47 %	65 bp	1,5	10,26 %
ML2	ML109	90,60 %	49 %	65 bp	4,2	19,9 %
ML1	ML110	78,50 %	45 %	65 bp	3,8	15,58 %
ML1	ML111	65,50 %	48 %	65 bp	0,8	7,97 %
ML1	ML112	75,00 %	46 %	65 bp	2,6	9,96
ML1	ML114	79,40 %	45 %	65 bp	3,4	14,6 %
ML2	ML115	89,40 %	49 %	65 bp	3,6	18,17 %

In der Abbildung 17 sind der Phred-Werte für alle Proben Basenposition dargestellt. Alle Proben liegen im grünen Bereich zwischen den Qualitätswert 30 und 35.

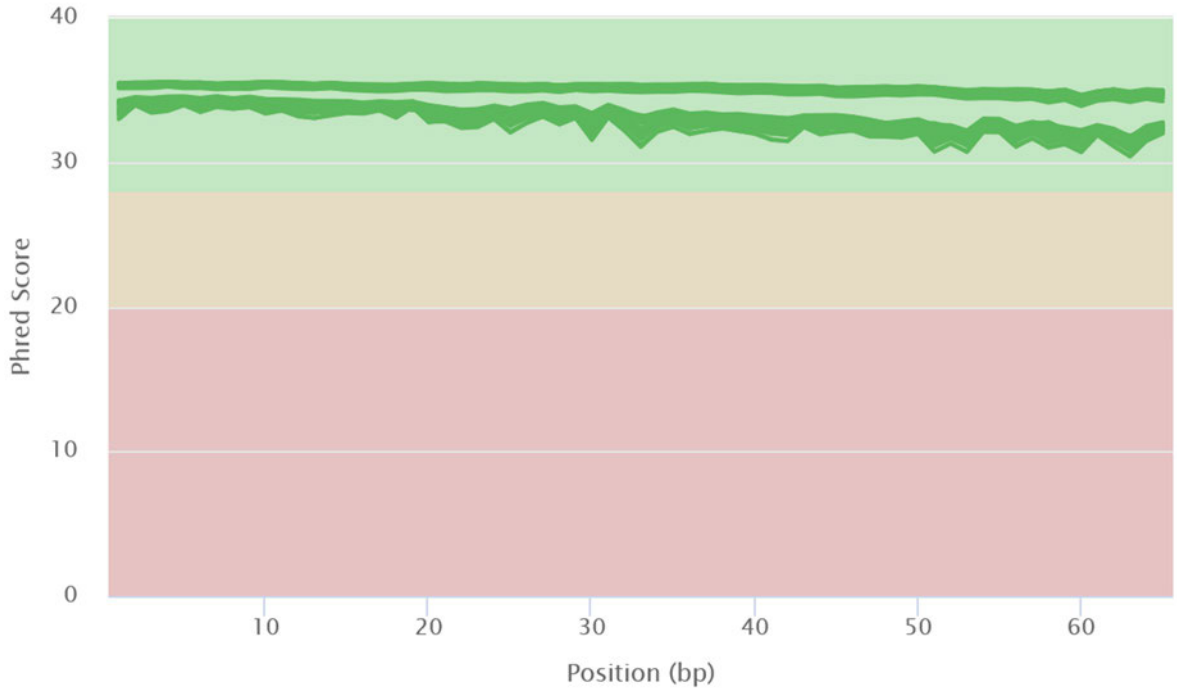

Abbildung 17: Der Phred-Qualitätswert der einzelnen Proben von ML1_R1 und ML2_R1 in Bezug auf die Basenposition.
In dieser Abbildung ist der Phred-Qualität-Wert über alle Basenpaare dargestellt. Die Werte von allen Proben liegen im grünen Bereich zwischen 30 und 35. Die Abbildung wurde vom Programm MULTIQC erstellt.

Die Abbildung 18 zeigt den Qualitätswert pro Sequenz der bearbeiteten Proben an. Der Verlauf der Kurve ist bei allen Proben ähnlich. Das Maximum liegt bei allen Proben bei einem Phred-Wert von 35. Die meisten Sequenzen die bei 35 liegt hat die Probe ML078 mit 6 mio. Reads. Der Kurvenverlauf steigt von etwa 20 bis 30 linear an und steigt ab 30 exponentiell bis 35 an. Bei allen Proben fällt der Qualitätswert wieder ab.

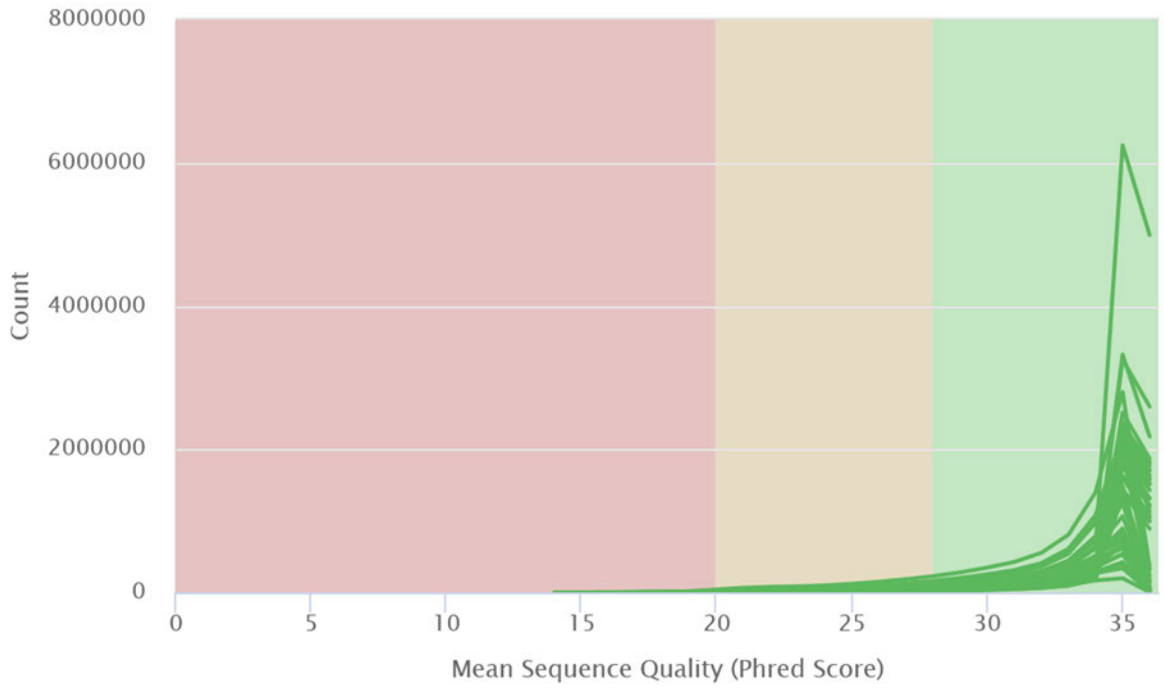

Abbildung 18: Der Phred-Qualitätswert pro Sequenzen der einzelnen Proben von ML1_R1 und ML2_R1.

Die Darstellung zeigt den Phred-Qualitäts-Wert über alle Sequenzen. Die Sequenzen liegen zum Größten Teil bei einem Wert zwischen 28 und 35. Die maximale Anzahl der Sequenzen liegt bei einem Wert von 35. Die Abbildung wurde von MULTIQC erstellt.

Die Abbildungen 19 und 20 zeigen für die Proben ML077 und ML023 die Basenverteilung über 65 bp an. Die beiden Verteilungen weisen wie die Rohdaten eine große Variabilität auf. Bei der Probe ML077 liegt die Basenverteilung bei der Position 37 bp bei T bei 44,8 %, C bei 18,7 %, A bei 10,5 %, G bei 26 %. Die Probe ML023 hat in der gleichen Basenposition 37 bp folgende Verteilung T liegt bei 43,5 %, C bei 19,9 %, A bei 17,8 %, G bei 18,9 %. Alle anderen Proben sehen ähnlich aus. Die ausgewählten Proben ML077 und ML023 sind nur Stichproben. Die meisten Proben haben Schwankungen bei der Basenverteilung von 10 % bis 50 %.

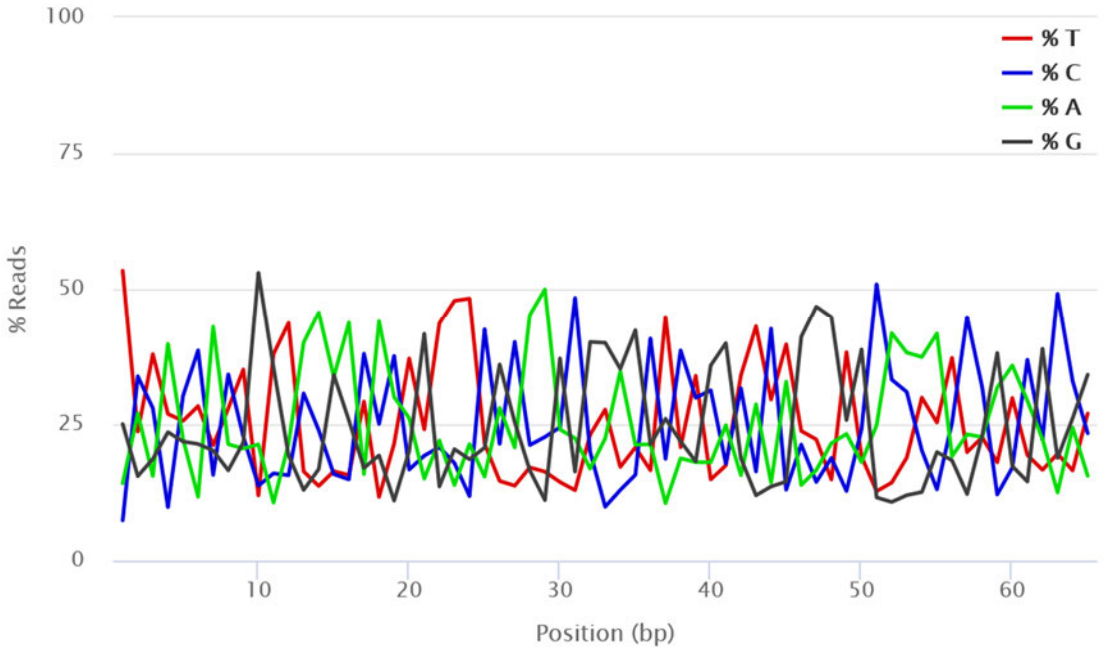

Abbildung 19: Basenverteilung der Reads der Probe ML077.
Die Kurven der Basen besitzen eine starke Variabilität und schwanken zwischen 10 % und 50 % der Reads. Die Abbildung wurde vom Programm MULTIQC erstellt.

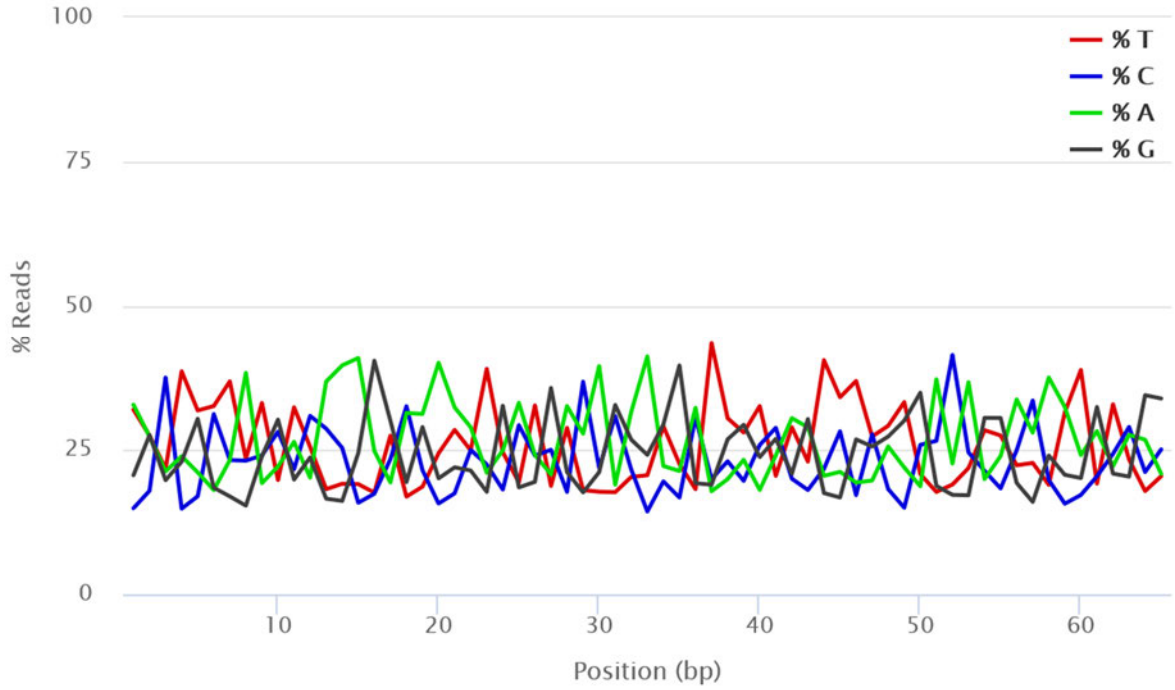

Abbildung 20: Basenverteilung der Reads der Probe ML023.
Die Kurven der Basen besitzen eine starke Variabilität und schwanken zwischen 10 % und 40 % der Reads. Die Abbildung wurde vom Programm MULTIQC erstellt.

Die Tabelle 18 gibt die Mittelwerte der Anzahl der Sequenzen und die Standardabweichung der beiden Bibliotheken ML1 und ML2 und die Zusammenführung ML1 + ML2. Die Werte wurden eingesetzt, um Proben die über- oder unterpräsentiert aus der Bibliothek zu entfernen. Der Mittelwert von ML1 liegt bei 3,13 M Sequenzen. Die Standardabweichung von ML1 ist bei 1,69. Damit wurde eine Wertebereich gebildet, der bei ML1 zwischen 1,43 M und 4,82 M Sequenzen liegt. Bei der zweiten Bibliothek ML1 liegt der Bereich von 2,69 M und 6,84 M Sequenzen. Die Zusammenführung ML1 + ML2 hat den Wertebereich zwischen 1,88 M und 6,01 M Sequenzen. Die Proben, die außerhalb des Wertebereichs lagen, wurden entfernt.

Tabelle 18: Mittelwerte der Anzahl der Sequenzen der Proben.
Angezeigt sind der Mittelwert und die Standardabweichung der Sequenzen der jeweiligen Proben der Bibliotheken. Aus der Standardabweichung wird ein Mengenbereich definiert, der es möglich macht, Proben, die unter- oder überpräsent sind, auszusortieren.

	ML1	ML2	ML1 + ML2
Mittelwert	3,13 M	4,77 M	3,95 M
Standardabweichung	± 1,69 M	± 2,08 M	± 2,06 M
Oberer Bereich	4,82 M	6,84 M	6,01 M
Unterer Bereich	1,43 M	2,69 M	1,88 M

Ausgang der Bearbeitung durch STACKS

Die Ausgabe der bearbeiteten und sortierten Rohdaten von *process_radtags* sind in der Tabelle 19 aufgeführt. Die Gesamtzahl der Sequenzen, die bearbeitet wurden, liegen bei ML1 bei 345 mio. Sequenzen, davon wurden 12 mio. Sequenzen durch das Trimmen der Adapter entfernt. 175 mio. Sequenzen wurden entfernt, weil der Barcode in den Sequenzen nicht gefunden worden ist. Ein geringer Teil der Sequenzen von 231,761 Reads sind wegen einem schlechten Qualitätswert entfernt worden. Nach den Abzügen blieb 45 % der Gesamtsequenzen von 157 mio. Reads übrig.

Tabelle 19: Ausgaben von *process_radtags.*
Dargestellt sind der Wert die nach der Bearbeitung von STACKS *process_radtags* ausgegeben werden. Dabei wird die Gesamtanzahl der Sequenzen gezeigt, welche Sequenzen Adapter-Sequenzen enthalten und entfernt werden, Sequenzen, bei der der Barcode nicht gefunden wurde und wieviel Sequenzen einen schlechte Qualitätswert haben. Am Ende steht die Anzahl der Reads, die für die weiter Bearbeitung verwendet werden.

	ML1	ML2
Totale Sequenzen	345.007.878	251.915.490
Reads die Adapter-Sequenzen enthalten	12.136.137	1.818.292
Barcode nicht gefunden	175.038.358	19.668.478
Schlechte Qualität	231.761	344.309
Beibehaltene Reads	157.601.622	230.084.411

Die zweite Bibliothek ML2 hat eine Gesamtzahl der Sequenzen von 252 mio. Reads. Nach den Abzügen durch das Adapter-Trimmen von 2 mio. Reads, durch die schlechte Qualität von 344.309 Seq. und von Reads, indem kein Barcode gefunden wurde von 20 mio., bleiben 230 mio. Reads übrig. Die beibehaltenen Sequenzen machen 91 % der Gesamtanzahl der Reads aus.

Zwei Werte aus der Ausgabe von dem Programmteil *populations* sind in Tabelle 20 dargestellt. Einmal die Anzahl der Loci und die Anzahl der Variantenstandorte die für die weitere Analyse mit STRUCTURE wichtig ist.

Tabelle 20: Ergebnis der Ausgabe des STACKS-Programm *populations.*
Ausgabe zeigt die Gesamtanzahl der gebildeten Loci und die Variantenstandort bzw. die Anzahl der Loci, die einen Unterschiede zwischen den Proben ab.

	ML1	**ML2**	**ML1 + ML2**
Anzahl der Loci	129.107	54.428	87.832
Variantenstandorte	14.434	12.535	21.233

Die größte Anzahl der Loci hat die erste Bibliothek ML1 mit 129.107 die Zusammenführen ML1 + ML2 hat 87.832 Loci und die zweite Bibliothek ML2 54.428 Locis. Bei der Anzahl der Variantenstandorte hat ML1 + ML2 die meisten mit 21.233, ML1 hat 14.434 und die zweite Bibliothek ML2 12.535.

Weiter Angaben von der Ausgabe von *populations* ist in Anhang A auf Englisch aufgeführt.

Populationsanalysedaten von STRUCTURE und STRUCTURE HARVESTER

Ergebnisse der ΔK-Bestimmung durch STRUCTURE HARVESTER

Die Abbildung von 21 bis 23 zeigt die Ergebnisse der ΔK Bestimmung an. Dabei sind die ΔK-Werte gegen die K-Werte aufgetragen.

Der ΔK für die erste Bibliothek ML1 hat bei einem Wert von 709,6 bei einen K-Wert von 3 ein Maximum. [Abb. 21] Die restlich vier K haben für ΔK einen Wert, der gegen Null geht.

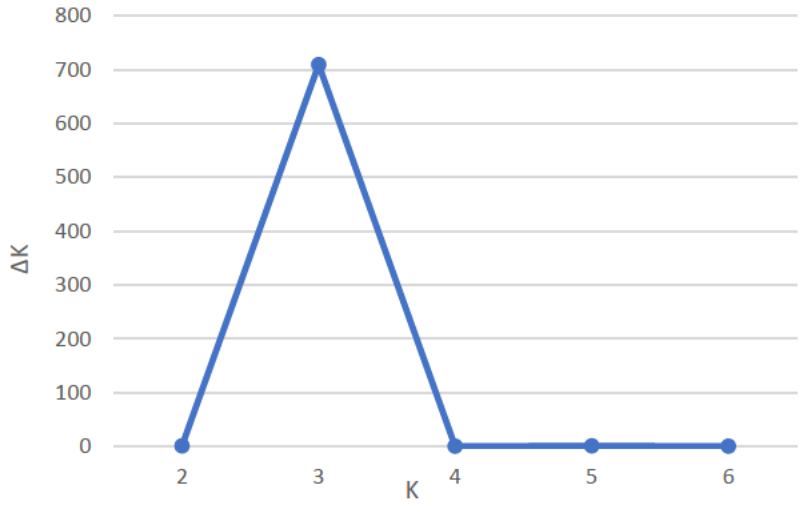

Abbildung 21: Bestimmung der Anzahl der stabilen Population ΔK von ML1.
In dem Diagramm ist ΔK gegen K aufgetragen. Die Kurve hat bei einen ΔK von 700 ein Maximum an der Stelle von K gleich 3.

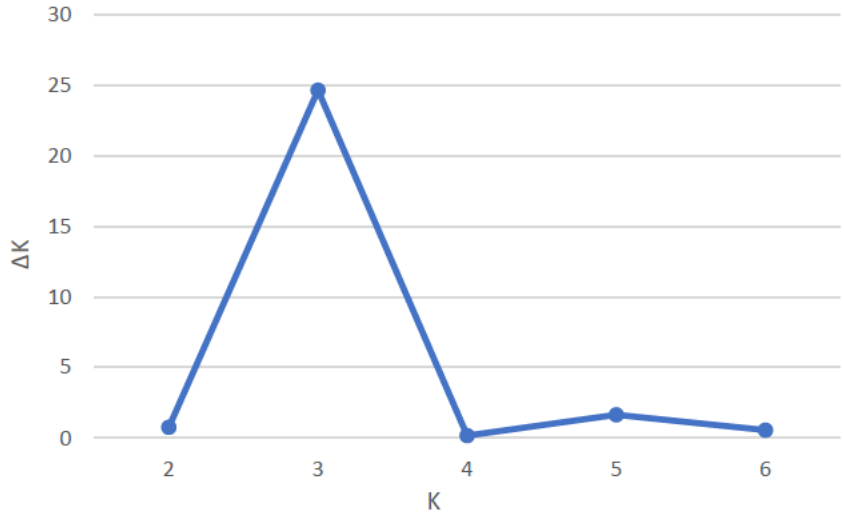

Abbildung 22: Bestimmung der Anzahl der stabilen Population ΔK von ML2.
In dem Diagramm ist ΔK gegen K aufgetragen. Die Kurve hat bei einen ΔK von 25 ein Maximum an der Stelle von 3 K.

Für den ΔK-Wert der zweiten Bibliothek ML2 hat die Kurve wie bei ML1 bei K = 3 ein Maximum. [Abb. 22} Der Wert von ΔK liegt an diesen Punkt bei 24,66. Bei einem Wert von K = 5 liegt ein kleines Maximum bei einen ΔK von 1,6. Die restlichen Punkte weisen einen Wert von ΔK unter 1 auf.

Die Abbildung 23 zeigt die Bestimmung von ΔK für die Zusammenführung von ML1 + ML2 an. Das Maximum liegt bei eine K-Wert von 4 bei einen ΔK von 5803,55. Die restlichen K-Werte zeigen keinen weiteren Anstieg in der Kurve an. [Abb. 23]

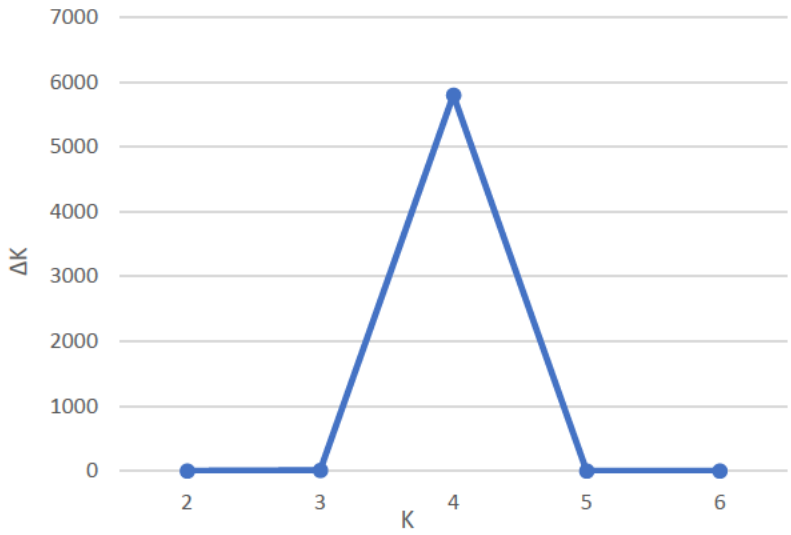

Abbildung 23: Bestimmung der Anzahl der stabilen Populationen ΔK von ML1 + ML2.
In dem Diagramm ist ΔK gegen K aufgetragen. Die Kurve hat bei einen ΔK von 6000 ein Maximum an der Stelle von 3 K.

Die Berechnungstabelle für die ΔK-Bestimmungen ist in Anhang B dargestellt. [Tab. 21 - 23]

Ergebnisse der STRUCTURE-Läufe mit dem stabilen K

In den Abbildungen von 24 bis 26 sind die bei den STRUCTURE-Läufen herstellten Q-Matrix grafisch dargestellt.

Für die erste Bibliothek ML1 wurde die Populationsstruktur mit einem K = 3 simuliert. Dabei stellt sich die in Abbildung 24 dargestellte Verteilung ein. Die K1-Population (blau) wird von den Proben aus DEUR gebildet, jeweils einen Wert von 1 nur einen kleinen Anteil finden sich in den Proben ML102 mit keiner Angabe vom Herkunftsort und ML089 aus den NLD. Die Proben aus HUN präsentieren die K2-Population (orange) mit einem Wert von 1. Von K2 habe folgenden Proben einen Anteil ML107, ML098, ML061, ML030 und ML023. Die K3-Population

(grau) dominiert die Proben aus NLD, DEUM, DEUK, aus SVK und die Probe ML102 ohne Angabe des Fundorts. Dabei habe die Proben ML104, ML090 und ML088 einen Wert von 1

Die zweite Bibliothek ML2 wurde mit einem K von 3 simuliert und die Verteilung ist in Abbildung 26 grafisch dargestellt. Die Variabilität der Population-Verteilung ist nicht stark ausgeprägt, alle Proben haben einen Wert von 1. Dabei gehören die Proben der K1-Population (blau) zu den Fundorten aus NLD und DEUR. Die K2-Population beschränkt sich nur auf die Proben aus DEUK. Die Proben aus der SVK und die Proben ohne Ortsangaben k. A. bilden die K3-Population (grau) ab.

Die Abbildung 25 zeigt den Simulationslauf der beiden zusammengeführten Bibliotheken ML1 + ML2 mit einen K von 4. Die blaue K1-Population dominiert die Proben aus DEUK mit einem Wert von 1. Weitere Anteile von K1 haben die Proben aus NDL, die Proben, die keine Angaben zum Fundort haben und die Probe ML104 aus DEUM. Die K2-Population (orange) hat ihren größten Anteil in den Proben aus SVK und ist Bestandteil der vier Proben ML101, ML100, ML099, ML102. Die Proben aus DEUR sind der F3-Population (grau) zu 100% zu geordnet. Die NLD-Proben weisen einen K3-Anteil von 20 % bis 60 % auf. Die Proben ML033, ML030, ML029, ML023 haben ein Anteil von 12 % bis 15 % von K3. Die K4-Populations (gelb) dominiert die Proben aus HUN und DEUM. Die Proben ML102 besitzt von K4 einen Anteil von 12 %. In der Probe ML102 sind alle Populationen vertreten, von die K2 mit 0,474 den größten Anteil ausmacht.

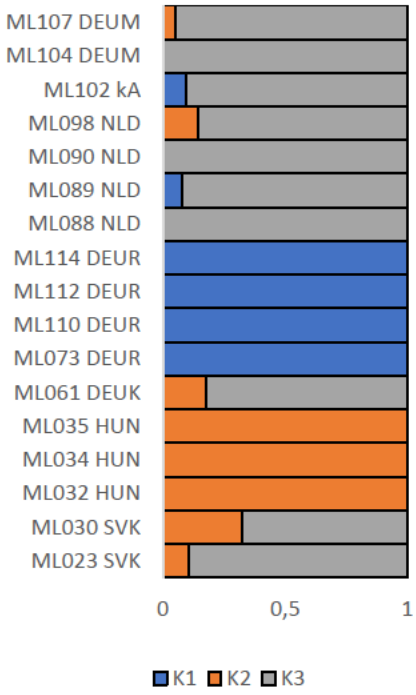

Abbildung 24: Populationsstruktur der Bibliothek ML1 bei eine K von 3.
Die Darstellung zeigt die Anteile der Populationen K von 0 bis 1 an die jede Probe eines bestimmten Fundortes hat.

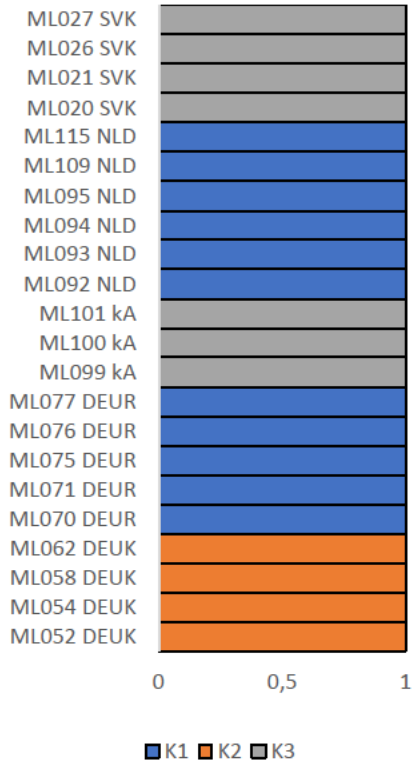

Abbildung 26: Populationsstruktur der Bibliothek ML2 bei einen K von 3.
Die Darstellung zeigt die Anteile der Populationen K von 0 bis 1 an die jede Probe eines bestimmten Fundortes hat.

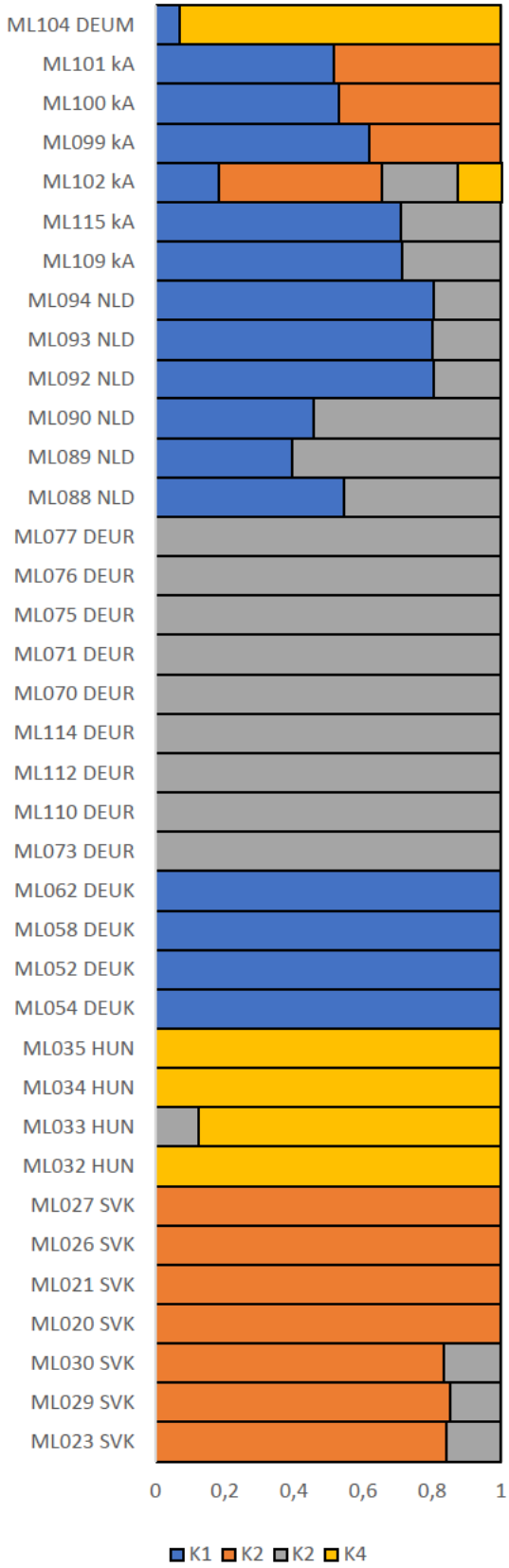

Abbildung 25: Populationsstruktur der Zusammengeführten Bibliotheken ML1 + ML2 bei einen K von 4
Die Darstellung zeigt die Anteile der Populationen K von 0 bis 1 an die jede Probe eines bestimmten Fundortes hat.

Diskussion

Einschätzung der Qualität der hergestellten ddRAD-Bibliothek

Die hergestellten Rohsequenzen der ddRAD-Bibliotheken ML1 und ML2 haben nach dem Phred-Qualitätswert eine gute Qualität. Der Wert lag, bei beiden, pro Basenposition über 30 [Abb. 13]. Da der Phred-Wert bei der Sequenzierung durch den Illumina-Sequenzierer erzeugt werden, gibt der Wert nur eine Aussage darüber, wie gut die DNA der Bibliothek sequenziert wurden und ob die Bibliotheken DNA enthält, die keine Verunreinigungen aufweist, die die Sequenzierung stört und somit die weitere Analyse verhindert. Der Phred-Wert gibt aber keine Aussage darüber, ob der Inhalt, z.B. ob die Basenfolge des Barcodes noch intakt ist. Da der Barcode ein wichtiges Element eine DNA-Bibliothek ist, wird auf die Beurteilung der Qualität die Basenverteilung pro Basenposition geschaut. Die beiden Bibliotheken ML1 und ML2 besitzen eine starke Variabilität zwischen den Basen, besonders die Unterschiede zwischen den komplementären Basen (A und T, C und G) von 10 % - 30 % bei einigen Basenpositionen fällt auf. [Abb. 15,16] Das allgemeine Verständnis ist, dass die komplementären Basen in ein Verhältnis von 1:1 zueinander sind und die Kurve deswegen parallel zueinander laufen sollten. Eine Erklärung könnte sein, dass bei der Sequenzierung ein Problem aufgetreten ist, oder im FASTQC-Manual steht dazu, das die Unterschiede durch überrepräsentierte Sequenzen wie z.B. Adapterdimere verursacht wurden. Für die ersten Bibliothek ML1 würde ich die überrepräsentierten Sequenzen ausschließen, weil der Anteil bei R1 bei 0% und bei R2 bei 3 % liegt. [Tab. 17]. Bei der zweiten Bibliothek ML2 liegen die überpräsenten Sequenzen weit höher bei 14 % (R1) und 21 % (R2) [Tab. 17], welches ein Grund für die hohe Variabilität der Basenverteilung bei ML2 sein kann. Ein weiterer Grund kann sein, dass die Phusion-Polymerase bei der PCR gestört wurde bzw. ungenau die Sequenzen verdoppelt hat. Die beiden Bibliotheken haben eine hohe PCR-Duplikation, ML2 liegt bei beide Läufen bei 93 % und bei ML1 liegen die Anteile der Duplikation bei 62,4 % (R1) und 73,3 % (R2). Die PCR verstärkt auch den Anteil an Adapterdimere, da die Primer an den Adapter anlagern. Eventuell kann es auch an der Qualität der Roh-DNA liegen, da die Proben nicht direkt vor der Herstellung der ddRAD-Bibliothek extrahiert worden sind, sondern für eine längere Zeit eingefroren waren.

Eine weitere Einschätzung, ob die Bibliothek eine gute Qualität hat, ist der Anteil der Sequenzen, bei denen der Barcode gefunden wird, gegenüber der Gesamtanzahl der

Sequenzen in den Rohdaten, da ohne Barcode keine Einteilung der Sequenzen zu den Proben erfolgen kann und keine Analyse oder Populationsuntersuchung stattfinden kann. Die Hälfte (50,7 %) der Sequenzen bei ML1 hat keinen Barcode. Bei der der zweiten Bibliothek ML2 liegt der Anteil von Sequenzen, bei den kein Barcode gefunden wurde, bei 7,8 %. [Tab. 20]. Die erste Bibliothek ML1 hat wesentlich weniger eingeteilte Sequenzen als ML2, obwohl ML1 in den Gesamtanzahl der Sequenzen 100 mio. Sequenzen mehr hat. [Tab. 20, 18]

Um die Qualität zu verbessern, wurde mit *process_radtags* die Adapter entfernt und die Reads mit schlechter Qualität (Phred-Wert 10) wurden entfernt. Dazu wurden die Readlänge auf 75 bp verändert, um die Verarbeitungsdauer der späteren Programme zu verkürzen, die Restadapter am Ende der Sequenzen zu entfernen und die überrepräsentierten Sequenzen entgegenzuwirken. CUTADAPT diente dazu, um die ersten 10 bp zu entfernen, um die starke Variabilität der Basenverteilung in diesen Bereich zu entfernen. Dabei muss aufgepasst werden, dass die Barcodes nicht mit abgeschnitten werden. Generell sollt beim Trimmen der Reads alle Basen entfernt werden, die durch die Herstellung der ddRAD-Bibliothek hinzugefügt werden, z.B. Adapter, Primer.

Die Qualität hat sich durch das Trimmen und das Entfernen der Reads mit schlechter Qualität verbessert, besonders bei den Proben von ML1 hat sich der Qualitätswert pro Sequenz mehr in den Bereich ab einen Wert von 28 verschoben, die Kurven sind abgeflachter [Abb. 18] als bei den Rohsequenzen von ML1. [Abb. 14] Die starke Variabilität der Basenverteilung in den ersten Basen wurde bei allen Proben entfernt, was zu einer Reduzierung der Unterschiede zwischen komplementäre Basen führte. Bei den Angaben zur Duplikation und überrepräsentierten Sequenzen hat sich nichts verändert. Die überrepräsentierten Sequenzen sind bei den Proben von ML1 im Gegensatz zu den Rohdaten angestiegen. Ein hoher Anteil von überrepräsentierten Reads kann auch biologische Gründe haben [FASTQC-Manual] z. B. eine bestimmte Sequenz ist in dieser Art dominant. Weiterhin kann sein, dass die Adapterdimere durch die starke Variabilität der Basenverteilung so verändert sind, dass sie mehr als eine Veränderung in den Basen hat, so dass *process_radtags* mit den Trimm-Adapter die Sequenzen nicht mehr erkennt.

Durch Veränderung der Parameter in den Programmen kann weiter die Qualität verbessert werden, dies war in dieser Bachelorarbeit zeittechnisch nicht möglich, da Programmen einiges an Zeit braucht, um die Daten zu berechnen.

Die Qualität der beiden ddRAD-Bibliotheken ML1 und ML2 würde ich als akzeptabel einschätzen, da es die ersten beiden Bibliotheken sind, die mit dem Protokoll hergestellt wurden und das nach den Schneiden und Reinigen der Proben mit *CUTADAPT* und *process_radtags* eine Qualitätsverbesserung eingetreten ist, besonders bei ML1, die schlechter war als ML2. Die Parameter-Duplikation und überrepräsentierte Reads, die nicht mit informatischen Mitteln behoben werden können und so viele Vermutungen zulassen, warum sie so hoch sind, würde ich für die Qualitätsbewertung außen vorlassen. Ein Parameter für Qualitätsbewertung muss eindeutiger sein

Erörterung der populationsgenetischen Analyse

Aus den Reads nach der Reinigung und Einteilen der Reads zu den Proben wurden bei ML1 157 Mio. Reads und bei ML2 230 Mio. Reads für die Herstellung eines Populationskatalog eingesetzt, um eine Populationsstruktur darzustellen. Dabei kamen für ML1 14.434 und für ML2 12.539 Variationsstellen heraus, in den sich die Proben gleichen oder unterscheiden. Eine Zusammenführung der beiden Bibliotheken nach Einteilung hat 21.233 Variationsstellen. Alle Läufe mit der STACKS-Pipeline hatten die gleichen Parametereinstellung. Durch die Herausnahme von Proben haben sich die Bibliotheken ML1 (17) und ML2 (22) in der Anzahl der Proben unterschieden. Des Weiteren unterscheiden sich die Bibliotheken in der Anzahl der Proben aus den verschiedenen Fundorten. ML1 enthält Proben aus 7 Fundorten, ML2 hat nur Proben aus 5 Fundorten. Es fehlen Proben aus Ungarn (HUN) und Munster (DEUM). Dadurch findet das Programm STACKS andere Verbindungen zwischen den Proben als bei der ersten Bibliothek ML1. Damit können die beiden Bibliotheken nicht miteinander verglichen werden. Deswegen wurden die Proben der beiden Bibliotheken zusammengeführt, um alle Proben miteinander zu vergleichen. Für alle drei Möglichkeiten wurde eine Populationsstruktur berechnet und geplottet [Abb. 24 - 26].

Durch eine ΔK-Bestimmung hat sich herausgestellt, dass die von STACKS erzeugten Datensätzen für die beiden Bibliotheken ML1 und ML2 eine stabile Struktur für 3 Populationen erzeugt. Die Zusammenführen ML1 + ML2 hat bei dem gleichen Parameter eine stabile Struktur bei 4 Population. Durch die Gründe, die oben angeführt worden sind und durch die stichpunktartige und gedächtnislose Berechnung (MCMC) die STRUCTURE benutzt, sind diese drei Populationsstrukturen untereinander nicht vergleichbar und stehen nur für sich.

Die berechneten Populationsstrukturen [Abb. 24 - 26] und die Bestimmung von ΔK bestätigt unsere Hypothese nur teilweise. Es habe sich aufgrund des isolierten Vorkommens der Heideschrecke nicht 6 -7 Populationen ausgebildet, sondern nur 3 bzw. 4. Das sich aber schon andere Populationen ausgebildet haben, zeigt schon, dass sich der Genotyp der Heideschrecke sich in den isolierten Gebieten verändert hat.

Weiterhin zeigen die Plots von ML1 und ML1 + ML2, dass einige Proben verschieden große Anteile von anderen Populationen tragen. Was für eine genetische Veränderung im Standort sprechen könnte oder eine Verbindung von zwei Populationen zeigt. Zum Beispiel: In der Abbildung der ML1 + ML2-Strucktur besteht die Proben aus NDL aus zwei Population einmal von DEUR (grau) und von DEUK (blau). [Abb.25] Es könnte auch ein Anzeichen für eine Wanderbewegung von Heideschrecken sein, die von DEUK über DEUR in die NDL weitergewandert sind.

Die Proben aus SVK und HUN stellen jeweils ihre eigene Population da. Anteile aus der SVK-Population befinden sich noch in vier Proben, die keine Abgaben zum Fundort haben. Eine Besonderheit sind die Proben ML102, die alle Population vereint hat und die einzelne Probe aus Munster (DEUM), die einen großen Anteil der Population aus HUN trägt, obwohl die beiden Fundorte am weitesten entfernt sind. Im Gegensatz dazu zeigt ML1 für die Proben aus DEUM, das sie eher eine Population mit den Proben aus NLD bildet.

Im Gegensatz zu den Strukturen von ML1 und ML1+ML2 weist die Struktur von ML2 überhaupt keine Variabilität innerhalb Proben bei den Anteilen der Population auf. Die Proben sind eindeutig einer Population zugeordnet. Die Proben von NLD und DEUR sind eine Population und die Proben, die keine Angabe des Fundortes haben, werden der Population der SVK-Proben zugeordnet. Die dritte Population sind die Proben aus DEUK.

Diese drei Beispiele sollten zeigen, dass jede Struktur eine eigene Analyse der Populationsverteilung zulässt, die von der einer anderen Struktur nicht bestätigt werden kann. Deswegen ist eine allgemeine Aussage über die Verteilung der Population der Heideschrecke in Mitteleuropa nicht möglich, weil jeder STRUCTURE-Lauf eine andere Verteilung ausgibt. Die einzige Aussage, die getroffen werden kann, dass die zwei ddRAD-Bibliotheken ML1 und ML2 und das informatische Zusammenführen der beiden zeigen, dass sich unterschiedliche Population gebildet haben, zwar 3 bis 4, aber keine Aussage getroffen werden kann, wie diese

verteilt sind. Um eine allgemeine Aussage zu tätigen, muss die Heideschrecke weiter untersucht werden, dies ist im Rahmen eine Bachelorarbeit jedoch nicht möglich ist.

Methodenbeurteilung und Ausblick

Die Ausgaben und Ergebnisse der Tapestation zeigen, dass mit den benutzen Protokoll zur Herstellung eine ddRAD-Bibliothek eine Bibliothek hergestellt werde kann, die für eine bioinformatische Populationsanalyse eingesetzt werden kann. In der Aufnahme der Tapestation von der ersten Bibliothek ML1 steigt der Peak bzw. Bande von 250 bp bis auf 400 bp an. Dieser Anstieg ist ein Anzeichen, das sich die Adapter, die für eine bioinformatische Unterscheidung der Proben wichtig sind, an die DNA-Fragmente gelagert haben. [Abb. 6 und 7] In der zweiten Bibliothek zeigt die Kurve ein ähnliches Bild, aber der Anstieg der Basenpaare beträgt nur etwa 70 bp. Der geringere Anstieg liegt an der durchgeführten Größenselektion bei ML1 die schon größeren Fragmente entfernt hat. [Abb. 9 und 10]. Die Bilder bzw. Kurve von den PCR-Pool von ML1 und ML2 zeigen nur geringe Veränderung in der Basenpaarlänge an. [Abb. 8 und 11] Die Verringerung der Basenpaare auf etwa 300 bp in den finalen Bibliotheken resultiert aus der Größenselektion mit dem Fragmentierer BluePippin der DNA-Fragmente von 300 bp selektiert hat. [Abb. 8 und 12]

Das Programm bzw. die Parameter sind so gewählt worden, um eine schnelle bioinformatische Analyse während der Bearbeitungszeit der Bachelorarbeit zu bekommen. Die Programme STACKS und STRCUTURE brauchen je nach Größe und Anzahl der Reads (STACKS, CUTADAPT) und Anzahl der Loci (STRUCTURE) unterschiedlich lange zum Rechnen, obwohl ein Großrechner eingesetzt wurde. Insbesondere das Programm STRUCTURE, das aus dem Jahr 2012 stammt und irgendwie nicht mit den neuen Versionen von JAVA auf den Lib Cluster und auf meinen PC unter UNIX/LINUX/WINDOWS nicht richtig lief, besonders das Laden/Speichern von Projekten in der WINDOWS-Version funktionierte nicht. Die Simulation liefen aber ohne Abstürze. Ich habe für die Versuche die Linux-Console-Version und die Windows-Version benutzt. Für eine weitere Analyse würde ich ein anderes Programm für die Populationsstruktur verwenden, besonders für eine tieferen Analyse der Populationsstruktur. STRUCTURE ist ein gutes Werkzeug, um schnell die ersten Informationen über eine Populationsstruktur zu erhalten. Anschließend kann mit diesen Daten die Analyse bzw. das weitere Vorgehen in einem Projekt oder einer Arbeit angepasst werden.

Zur Verbesserung der Qualität bzw. um PCR-Duplikation, überrepräsentierte Sequenzen und starke Variabilität der Basen zu verhindern, sollten die Extraktion direkt vor der Herstellung der DNA-Bibliothek stattfinden, damit die DNA durch die Langzeitlagerung nicht beschädigt wird. Um die Adapterdimere zu verhindern, sollte die eingesetzte Menge der Adapter reduziert werden bzw. an die Menge der eingesetzten Proben-DNA gekoppelt werden. Da bei den Bibliotheken die PCR-Duplikation hoch war, würde ich die Zyklen des PCR-Laufes reduziert, etwa um 2 bis 5 Zyklen. Generell sollte die Herstellung mit frischen Materialen und Chemikalien erfolgen. Bei diesen Bibliotheken wurden Adapter eingesetzt, die 4 Jahre alt waren. Für die Enzyme sollte die Angabe der Firma wie die Zugabe-Volumen und die Inkubationszeit im Protokoll angepasst werden. Da die Bibliotheken eine hohe Anzahl an Reads hatten, bei wenigen Proben, sollte die Menge der eingesetzten DNA reduziert werden und abhängig von der Anzahl der eingesetzten Proben sein. Eine DNA-Menge von 200 ng bis 750 ng sollte ausreichen. Eine Reduzierung der DNA-Menge würde auch zu Verringerung von Kontamination von Dimeren und Artefakte führen.

Um die Information über die Populationsverteilung der Heideschrecke in einer ddRAD-Bibliotheken zu verbessern, müsst die Anzahl der DNA-Proben aus den Fundorten erhöht werden. Die Anzahl der Proben pro Fundort muss gleich sein, damit kein Fundort in der Analyse bevorzugt wird. Das war bei den hergestellten Bibliotheken nicht der Fall. Zum Beispiel waren in ML1 zwei DEUM-Proben und die restlichen Fundorte hatten teilweise mindestens 3 Proben. Eventuelle könnte noch eine Probe die artenverwandter ist oder eine fernverwandte Art der Heideschrecke hinzugefügt werden, um die Analyse-Methoden zu kontrollieren, denn zwischen denen und den betrachten Proben muss es Unterschiede geben.

Da es sich bei der bioinformatischen Analyse eine Population, um statistische und stochastische Methoden handeln, hilft es, wenn mehr Proben und somit mehr Information eingesetzt werden. Je mehr Information das stochastische System bekommt, umso weniger Fehler machen die Programme z.B. STRUCTURE, um eine Populationsstruktur zu simulieren. Der MCMC-Algorithmus ist sehr variabel, wenn nur kleine Zeitschritte gewählt werden, da es lange braucht, damit MCMC einen stationären Bereich erreicht. indem nur geringe Schwankung der Ergebnisse auftauchen. In dieser Arbeit wurde aus Zeitgründen nur Läufe im unteren Bereich der Simulationszeit durchgeführt, besonders bei der Bestimmung von ΔK für die burnin-Zeit, die bei 5000 lag. Diese burnin-Zeit kann erhöht werden, um ein stabileres

Plateau von MCMC zu erreichen, um für den restlichen Verlauf der Simulation einen geringeren Fehler zu haben.

Das Programm STACKS kann mit dem Variieren seine Parameter (-m, -M, -n) die Bibliothek weiter verbessern. Dies ist in dieser Arbeit nicht passiert, weil die Zeit gefehlt hat, um den besten Wert für die Parameter zu finden. Für die Arbeit wurden die Standardeinstellung von -m (Mindestabdeckungstiefe, die für eine Stacks-Bildung benötig wird, Standard: 3) und -M (maximaler zulässiger Abstand der Nukleoiden zwischen den Stacks, Standard: 2) gewählt und -n (erlaubte Unterschiede zwischen den Loci der Proben, um ein Katalog zu erstellen, Standard: 1) mit zwei festgelegt.

Die bioinformatische Analyse wurde so konzipiert, dass es als einer schnellen Analyse dient, um einen ersten Überblick über die Populationsstruktur eine Art zu bekommen.

Danksagung

Ich bedanke mich bei Dr. Oliver Hawlitschek und den Leibnitz-Institut zur Analyse des Biodiversitätswandels, weil Sie es mir möglich gemacht haben, dass ich eine interessanten und abwechslungsreichen Bachelorarbeit schreiben konnte. Des Weiteren bedanke ich mich bei meiner Familie, die mich als langjährigen Studenten ausgehalten habe und mich immer unterstützt haben. Weiter hin bedanke ich mich nochmal bei Oliver für seine Tipps und die Tätigkeit als zweiter Gutachter. Ein weiterer Dank geht an Herrn Prof. Dr. Béthune, der mein erster Gutachter ist und mich auch schon in meinem Praxissemester betreut hat.

Quellen

Andrews, S. (2010). FastQC. *Babraham Bioinformatics*.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11).

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: Building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, *1*(3).

Catchen, J. M., Cresko, W. a, Hohenlohe, P. a, Amores, A., & Bassham, S. (2016). Stacks Manual. *Most*.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, *14*(8).

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, *8*(3).

Hawlitschek, Oliver, (2019). Protocol for the computational analysis of raw data from ddRAD sequencing on Illumina MiSeq using STACKS.

Knoop, Volker, Müller Kai, (2009), Gene und Stammbaume: Ein Handbuch zur molekularen Phylogenetik, 2. Auflage, Spektrum Akademischer Verlag.

Pappas, F., & Palaiokostas, C. (2021). Genotyping strategies using ddrad sequencing in farmed arctic charr (Salvelinus alpinus). *Animals*, *11*(3).

Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution*, *8*(10).

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012a). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5).

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012b). protocol to ddRAD. *PLoS ONE*, *7*(5).

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2).

Pritchard, J. K., Xiaoquan Wen, Daniel Falush, (2010). Documentation for STRCUTURE software: Version 2.3.

Reinard, Thomas, (2010/2021). Molekularbiologische Methoden 2.0, 3. Auflage, utb.

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21).

Yan, J., Vétek, G., Pal, C., Zhang, J., Gmati, R., Fan, Q. H., Gunawardana, D. N., Burne, A., Anderson, D., Balan, R. K., George, S., Farkas, P., & Li, D. (2021). ddRAD sequencing: an emerging technology added to the biosecurity toolbox for tracing the origin of brown marmorated stink bug, Halyomorpha halys (Hemiptera: Pentatomidae). *BMC Genomics*, *22*(1).

www.orthoptera.ch/arten/item/gampsocleis-glabra, Datum: 08.10.2022 Uhr: 14:51

www.genome.gov/genetics-glossary/Contig, Datum: 30.11.2022, Uhr: 21:07

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang

Anhang A*: populations*-Ausgabe

ML1 *populations*-Ausgabe

Removed 47311 loci that did not pass sample/population constraints from 176418 loci.
Kept 129107 loci, composed of 16570540 sites; 3420 of those sites were filtered, 14434 variant sites remained.
Number of loci with PE contig: 129024.00 (99.9%);
Mean length of loci: 118.39bp (stderr 0.07);
Number of loci with SE/PE overlap: 40876.00 (31.7%);
Mean length of overlapping loci: 98.70bp (stderr 0.06); mean overlap: 23.80bp (stderr 0.03);
Mean genotyped sites per locus: 121.39bp (stderr 0.06).

ML1 *populations*-Ausgabe

Removed 71588 loci that did not pass sample/population constraints from 126016 loci.
Kept 54428 loci, composed of 6856822 sites; 8664 of those sites were filtered, 12535 variant sites remained.
Number of loci with PE contig: 54428.00 (100.0%);
Mean length of loci: 115.98bp (stderr 0.14);
Number of loci with SE/PE overlap: 28893.00 (53.1%);
Mean length of overlapping loci: 111.34bp (stderr 0.16); mean overlap: 22.45bp (stderr 0.04);
Mean genotyped sites per locus: 121.27bp (stderr 0.13).

ML2 + ML1 *populations*-Ausgabe

Removed 153278 loci that did not pass sample/population constraints from 241110 loci.
Kept 87832 loci, composed of 11410404 sites; 14709 of those sites were filtered, 21233 variant sites remained.
Number of loci with PE contig: 87825.00 (100.0%);
Mean length of loci: 119.92bp (stderr 0.10);
Number of loci with SE/PE overlap: 33786.00 (38.5%);
Mean length of overlapping loci: 108.25bp (stderr 0.12); mean overlap: 23.02bp (stderr 0.03);
Mean genotyped sites per locus: 123.69bp (stderr 0.09)

Anhang B: ΔK-Bestimmung

Tabelle 21: ΔK-Bestimmung von ML1.

Berechnungstabelle für die Abbildung 21.

K	Wiederholung	Mittelwert Ln(P(K))	Stdev Ln(P(K))	L´(K)	L´´(K)	ΔK
1	3	-389889,03	41,10	kA.	kA.	kA.
2	3	-468060,97	173580,06	-78171,93	188633,03	1,09
3	3	-357599,87	9985,79	110461,10	7085944,63	709,60
4	3	-7333083,40	12115554,07	-6975483,53	5795631,73	0,48
5	3	-8512935,20	7049564,19	-1179851,80	5182318,30	0,74
6	3	-14875105,30	16621565,62	-6362170,10	10171581,03	0,61
7	3	-11065694,37	7785020,25	3809410,93	kA.	kA.

Tabelle 22: ΔK-Bestimmung von ML2.

Berechnungstabelle für die Abbildung 22.

K	Wiederholung	Mittelwert Ln(P(K))	Stdev Ln(P(K))	L´(K)	L´´(K)	ΔK
1	3	-409207,50	67,71	kA.	kA.	kA.
2	3	-384514,00	6098,85	24693,50	4705,40	0,77
3	3	-355115,10	444,30	29398,90	10956,80	24,66
4	3	-336673,00	5635,46	18442,10	980,27	0,17
5	3	-319211,17	10657,78	17461,83	17445,77	1,64
6	3	-319195,10	10938,16	16,07	6043,60	0,55
7	3	-325222,63	10269,77	-6027,53	kA.	kA.

Tabelle 23: ΔK-Bestimmung von ML1 + ML2.

Berechnungstabelle für die Abbildung 23.

K	Wiederholung	Mittelwert Ln(P(K))	Stdev Ln(P(K))	L´(K)	L´´(K)	ΔK
1	3	-904807,70	322,39	kA.	kA.	kA.
2	3	-836964,30	383,47	67843,40	480,80	1,25
3	3	-769601,70	2008,69	67362,60	19749,23	9,83
4	3	-721988,33	4932,65	47613,37	28626865,47	5803,55
5	3	-29301240,43	15590571,69	-28579252,10	5541114,87	0,36
6	3	-63421607,40	108643101,43	-34120366,97	96852241,77	0,89
7	3	-689732,60	7331,21	62731874,80	kA.	kA.

Eidesstattliche Erklärung

Ich versichere hiermit, dass ich die vorliegende Bachelorarbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommenen Stellen sind unter Angabe der Quellen kenntlich gemacht.

Ort, Datum:

_____

Carsten Bruns