Eingehende Untersuchung des signifikanten quantitativen trait locus assoziiert mit Kalbemerkmalen auf Chromosom 18 in Holstein-Friesian-Rindern

von Nina Teresa Dachs

Inaugural-Dissertation zur Erlangung der Doktorwürde
der Tierärztlichen Fakultät der Ludwig-Maximilians-Universität
München

Eingehende Untersuchung des signifikanten quantitativen trait locus assoziiert mit Kalbemerkmalen auf Chromosom 18 in Holstein-Friesian-Rindern

von
Nina Teresa Dachs
aus Starnberg

München 2021

Aus dem Veterinärwissenschaftlichen Department
der Tierärztlichen Fakultät
der Ludwig-Maximilians-Universität München

Lehrstuhl für Molekulare Tierzucht und Biotechnologie

Arbeit angefertigt unter der Leitung von:
Priv.-Doz. Dr. Ivica Međugorac,
Arbeitsgruppe Populationsgenomik

Mitbetreuung durch:
Dr. Elisabeth Hannemann und
Ph.D. Maulik Upadhyay,
Arbeitsgruppe Populationsgenomik

Gedruckt mit der Genehmigung der Tierärztlichen Fakultät
der Ludwig-Maximilians-Universität München

Dekan: Univ.-Prof. Dr. Reinhard K. Straubinger, Ph.D.

Berichterstatter: Priv.-Doz. Dr. Ivica Međugorac

Korreferent/en: Prof. Dr. Armin M. Scholz

Tag der Promotion: 17. Juli 2021

Für Levi

INHALTSVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

ANN	Artificial Neural Network
BAM	Binary Alignment Map
BGVD	Bovine Genome Variation Database and Selective Signatures
BLUP	Best Linear Unbiased Prediction
bp	Basenpaar(e)
BTA	*Bos taurus* Autosom
BWA	*BURROWS-WHEELER ALIGNER*
bzw.	beziehungsweise
°C	Grad Celsius
ca.	circa
CCR	Konzeptionsrate
Cl^-	Chloridionen
cLDLA	combined Linkage Disequilibrium Linkage Analysis
cm	Zentimeter
cM	Centimorgan
DH	Deutsche Holstein
d.h.	das heißt
DGVa	Database of Genomic Variants Archive (Ensembl)
dGW	direkte genomische Zuchtwerte
DMR	Differentiell methylierte Region
DNA	Desoxyribonukleinsäure
DP	Depth of Coverage
Dr.	Doktor
D_{RM}	Diplotype Relationship Matrix
DSB	Doppelstrangbruch
DSD	Disorder of Sexual Development
e.V.	eingetragener Verein
engl.	englisch
etc.	et cetera
FDR	False Discovery Rate
FV	Fleckvieh
GATK	*GENOME ANALYSIS TOOLKIT*

Gbp	Gigabasen (1 Gbp entspricht 1.000.000.000 bp)
GPU	Graphical Processing Units
GQ	Genotypenqualität
G_{RM}	Additive Genotype Relationship Matrix
GWAS	Genomweite Assoziationsstudie
gZW	Genomisch unterstützter Zuchtwert
HC	*HaplotypeCaller*
HD	High Density
HF	Holstein-Friesian
HMM	Hidden Markov Model
HR	Homologe Rekombination
i.d.R.	in der Regel
IBD	Identity by Descent
IBS	Identity by State
IGV	*Integrative Genomics Viewer*
Inc.	Incorporated
Indel	Insertion und Deletion
KBV	Kärntner Blondvieh
Kb	Kilobasen (1 kb entspricht 1.000 bp)
kg	Kilogramm
KI	Konfidenzintervall
LAFUGA	Laboratory of Functional Genome Analysis
LCR	Low Copy Repeat
LD	Linkage Disequilibrium
LOD	Logarithm of the Odds
LRT	Likelihood Ratio Teststatistik
LS	Lernstichprobe
Lt	Lebenstag
µg	Mikrogramm
m/p KV	maternaler/paternaler Kalbeverlauf
m/p TG	maternale/paternale Totgeburt
MAF	Minor Allele Frequency
Mbp	Megabasen (1 Mbp entspricht 1.000.000 bp)
Na^+	Natriumionen
NAHR	Nicht-allelische homologe Rekombination
NCBI	National Center for Biotechnology Information

ng	Nanogramm
NGS	Next-Generation Sequenzierung
NHEJ	Non-homologous End Joining
NR56	Non-Return-Rate 56
ONT	Oxford Nanopore Technologie
PacBio	Pacific Biosciences
Ph.D	Philosophical Doctorate
POA	Partial Order Alignment
Prof.	Professor
PS	Phred-Score
QTL	Quantitativer Trait Locus
RNA	Ribonukleinsäure
RZ	Rastzeit
RZG	Relativer Gesamtzuchtwert
SAM	Sequence Alignment Map
SD	Segmentale Duplikation
Sg	genetische Standardabweichung
SMRT	Single-molecule real-time Sequencing
SNP	Single Nucleotide Polymorphism
SRA	Sequence Read Archive
SV	strukturelle Variante
u.a.	unter anderem
UAR	Unified Additive Relationships
usw.	und so weiter
v.a.	vor allem
VCF	Variant Calling Format
VIT	Vereinigte Informationssysteme Tierhaltung w.V.
VZ	Verzögerungszeit
WGS	Whole-genome sequencing
w.V.	wirtschaftlicher Verein
z.B.	zum Beispiel
ZW	Zuchtwerte
ZWS	Zuchtwertschätzung

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

ANHANGSÜBERSICHT

1. Einleitung

Das Holstein-Friesian (HF) Rind ist mit einer durchschnittlichen Milchleistung von ca. 9.200 kg Milch pro Jahr (WHFF, 2018) eine der bedeutendsten Milchviehrassen weltweit und repräsentiert darüber hinaus mit ca. 53 % den Großteil der gehaltenen Rinderrassen in Deutschland (DEUTSCHER HOLSTEIN VERBAND, 2019). Durch die stetig steigende Nachfrage an tierischen Produkten im In- und Ausland verzeichnete die Bundesrepublik Deutschland im Jahr 2020 einen Produktionswert der tierischen Erzeugnisse und Tierproduktion von 26,3 Milliarden Euro, wovon 10,8 Milliarden Euro auf Milch und Milcherzeugnisse aller milchgebenden Spezies entfielen. Insgesamt wurden dabei in Deutschland im Jahr 2020 ca. 32 Millionen Tonnen Milch aller milchproduzierenden Tierspezies verarbeitet (BUNDESANSTALT FÜR LANDWIRTSCHAFT UND ERNÄHRUNG, 2020). Ein solch hohes Produktionsniveau benötigt eine robuste Tierpopulation, die durch stetig wachsende Kenntnisse in der Tierzucht und Genetik im Rahmen der Zuchtziele nachhaltig gesund und leistungsfähig erhalten wird. Durch Inzuchtdepression, welche sich unter anderem durch eine sinkende Fruchtbarkeit und steigende Schwer- und Totgeburtenraten in milchleistungsstarken Rassen manifestiert, entstehen jedoch immense ethische als auch wirtschaftliche Problem. Wie aus dem Jahresbericht 2019 des Vereinigten Informationssystem Tierhaltung (VIT) in Verden hervorgeht, war die Hauptabgangsursache in Rinderbetrieben unabhängig von der gehaltenen Rasse mit 19,4 % Unfruchtbarkeit (VEREINIGTE INFORMATIONSSYSTEME TIERHALTUNG W.V., 2019). Fruchtbarkeitsparameter sowie der Verlauf einer Kalbung werden neben Haltungs- und Managemententscheidungen zusätzlich von genetischen Faktoren wesentlich beeinflusst. Vor mehr als 20 Jahren wurde ein Quantitativer Trait Locus (QTL) mit hoch signifikanten Effekten auf Fruchtbarkeits- und Körperkonstitutionsmerkmale auf dem *Bos taurus* Autosom 18 (BTA18) kartiert. Dieser QTL konnte bisher ausschließlich in reinrassigen Holstein und mit Holstein veredelten Kreuzungslinien nachgewiesen werden. Durch diverse Studien gelang es, den QTL zwischen 50 und 60 Mbp auf BTA18 zu lokalisieren und eine Assoziation auf die Merkmale Fertilität, Kalbeverlauf, Totgeburt, Trächtigkeitsdauer und verschiedenen Merkmale der Körperkonstitution einzugrenzen (COLE et al.,

2009; SAHANA et al., 2011; MAO et al., 2016; MÜLLER et al., 2017; FANG et al., 2019; PURFIELD et al., 2020). Sowohl der Marker rs109478645 an der Position 57.137.302 bp des bovinen Referenzgenoms ARS-UCD1.2 (COLE et al., 2009) als auch der Marker rs381577268 bei 57.816.137 bp (FANG et al., 2019; PURFIELD et al., 2020) wiesen eine signifikante Assoziation mit diesem QTL auf. Darüber hinaus gelang es zwei vielversprechende Kandidatengene zu identifizieren, darunter das *sialic acid binding IG-like lectin* (*SIGLEC*) *13* Gen (57.136.157 – 57.142.779 bp) (COLE et al., 2009) und das *Zink Finger Protein 613* (*ZNF613*) Gen (57.774.874 – 57.816.078 bp) (FANG et al., 2019). In einem vorherigen Projekt der Arbeitsgruppe Populationsgenomik (MÜLLER et al., 2017) der Tierärztlichen Fakultät der LMU München konnte zudem ein Haplotyp lokalisiert werden, welcher in einem populationsweitem Kopplungsungleichgewicht mit den Merkmalen paternaler Kalbeverlauf und Totgeburt in Deutschen Holsteins (DH) steht. Der detektierte kausale Haplotyp reichte von 57.941.736 – 58.442.683 bp und erklärte vollständig die beobachteten QTL-Effekte der durchgeführten Kartierungsstudie. Darüber hinaus konnte nachgewiesen werden, dass der vermutlich kausale Haplotyp mit einer Frequenz von 13,3 % in der DH Population auftritt. Das bedeutet bei einer Population von 22,7 Millionen Holstein Tieren in der Europäischen Union und den USA (WHFF, 2018) wären 3,01 Millionen Holstein Tiere Träger des hoch signifikanten QTL mit dem größten negativen Einfluss auf Wirtschaftlichkeit, Tiergesundgeit und Tierwohl.

Trotz der enormen ethischen und wirtschaftlichen Problematik dieses QTL gelang es bisher nicht die kausalen Mutationen und/oder eindeutige Kandidatengene zu entschlüsseln. Daher war das Ziel dieser Arbeit, eine erneute Untersuchung der Kandidatenregion auf BTA18 mittels kombinierter Kopplungsungleichgewichts- und Kopplungsanalyse (cLDLA) unter Verwendung des bovinen Referenzgenoms ARS-UCD1.2 durchzuführen. Zu diesem Zweck wurde das Studiendesign von MÜLLER et al. (2017) verwendet und an die neuen Bedingungen dieser Studie angepasst. Darüber hinaus wurden Oxford Nanopore Long-Read Sequenzierungen und *de novo* Assembly Techniken angewendet, um die Region des QTL genauer zu untersuchen. Zusätzlich kamen verschiedenste Analysen zur Identifikation von Einzelnukleotid-Polymorphismen (SNPs), strukturellen Varianten (z.B. Insertion und Deletion) und segmentalen Duplikationen zum Einsatz.

2. Literaturübersicht

2.1. Das Deutsche Holstein Rind

2.1.1. Zuchtziele

Die derzeitigen Zuchtziele des Bundesverbandes Rind und Schwein e.V. sehen für das das Deutsche Holstein (DH) eine leistungsstarke, gesunde und langlebige Kuh vom milchbetonten Typ vor. Eine hohe Tagesleistung über viele Laktationen hinweg wird unter anderem durch ein widerstandsfähiges Fundament sowie ein gesundes und gut melkbares Euter erreicht. Zusätzlich sollte das Deutsche Holstein über ein hohes Futteraufnahmevermögen bei einer ausgezeichneten Futterverwertung verfügen. Das genetische Potenzial für den Komplex Milchleistung ermöglicht eine Leistung von über 10.000 kg Milch, in einem Zeitraum von 305 Tagen Leistung, mit einem Fettanteil von 4 % und einem Proteinanteil von 3,5 %. Folglich kann eine starke Lebensleistung von 40.000 kg Milch pro Kuh realisiert werden. Darüber hinaus zeichnet sich eine rentable Holstein Kuh durch eine gute Fruchtbarkeit und komplikationslose Kalbungen bei einem Erstkalbealter zwischen 25 und 28 Monaten aus. Das äußere Erscheinungsbild wird durch eine Kreuzhöhe von 145 – 156 cm und einem Gewicht von 650 – 750 kg bestimmt (BUNDESVERBAND RIND UND SCHWEIN, 2021).

2.1.2. Zuchtwertschätzung

Um die Zuchtziele einer Rinderrasse zu erreichen bzw. diese immer weiter zu optimieren, werden mit Hilfe der Zuchtwertschätzung jene Tiere ausgewählt, die für die Weiterzucht am besten geeignet sind. Durch Selektion der Elterntiere anhand ihrer Zuchtwerte, kann ein Zuchtfortschritt in der nächsten Generation erwartet werden. Das bedeutet, dass mit dem Zuchtwert nicht die eigene Leistung eines Tieres bewertet wird, sondern die Ausprägung der vererbten Merkmale in der nachkommenden Generation (FÜRST, 2021). Das Vereinigte Informationssysteme Tierhaltung (VIT) in Verden ist für die überregionale Zuchtwertschätzung (ZWS) der Rassen Deutsche Holsteins mit den Farbschlägen Schwarzbunt und Rotbunt, Rotvieh/Angler, Jersey und Deutscher Schwarzbunter Niederungsrinder verantwortlich. Daten für

Milchleistungs- und Zuchtleistungsmerkmale sowie Zellzahl stammen sowohl aus den Kontrollverbänden Deutschlands als auch aus Österreich. Seit August 2010 werden neben den klassischen Zuchtwerten (ZW), welche mit Hilfe von Eigen- und Nachkommenleistung ermittelt werden, auch genomische Informationen über den direkten genomischen Zuchtwert (dGW) herangezogen (VIT, 2021). Zur Berechnung des direkten genomischen Zuchtwerts ist eine sogenannte Lernstichprobe notwendig. Diese umfasst alle Tiere, die einerseits genotypisiert wurden und zu denen andererseits klassische Leistungsdaten vorliegen. Zur Schätzung des genomischen Zuchtwerts müssen außerdem alle genetischen Marker (SNPs) bekannt sein, die mit einem Merkmal in Verbindung stehen. Um nun die genomische Zuchtwertformel eines Merkmals zu erhalten, werden die SNP-Genotypen mit den klassischen Leistungsdaten der Lernstichproben-Tiere verglichen. Anhand einer Summenformel können dann den SNP-Effekten zugeordnete Zahlenwerte, welche auf der Merkmalsausprägung eines Tieres basieren, zum direkten genomischen Zuchtwert zusammengezählt werden. Die Sicherheit der direkten genomischen Zuchtwertformel hängt daher von der Größe der Lernstichprobe und der Sicherheit der klassisch geschätzten Leistungsdaten ab (VIT, 2021). Mit Hilfe dieser Schätzmethode ist es darüber hinaus möglich, die genomischen Zuchtwerte von Jungbullen über deren Vorfahren zu ermitteln, unabhängig davon, ob für diese Tiere Eigen- und Nachkommenleistungen zur Verfügung stehen (FÜRST, 2021). Hierbei wird angenommen, dass verwandte Tiere mit identischen Marker-Allelen an einer bestimmten Position die gleichen Effekte auf das Merkmal haben (OLSEN et al., 2004). Stehen für ein Tier jedoch sowohl Daten der Eigen- und Nachkommenleistung als auch der direkte genomische Zuchtwert zur Verfügung, werden diese in Form des genomisch unterstützten Zuchtwerts (gZW) kombiniert veröffentlicht (VIT, 2021).

Limitierender Faktor der Sicherheit der genomischen Zuchtwertschätzung ist die Heritabilität eines Merkmales. Die Heritabilität (h^2) oder Erblichkeit wird zwischen 0 und 1 angegeben, wobei 1 bedeutet, dass ein Merkmal allein von den additiv-genetischen Effekten abhängt. Weist ein Merkmal stattdessen nur eine geringe Heritabilität auf, ist dieses Merkmal nur zum Teil durch additiv-genetische Effekte bedingt. Der Einfluss nicht-additiv-genetischer Effekte wird durch genetische Interaktionen wie Dominanz und Epistasie bestimmt. Darüber hinaus werden Merkmale zusätzlich von verschiedensten

Umwelteffekten (z.B. Management, Haltung und Fütterung) beeinflusst. Beim Rind sind die Merkmale mit der niedrigsten Heritabilität Fruchtbarkeit (0,02, d.h. nur 2 % additiv-genetisch bedingt), Kalbeverlauf (0,05) und Totgeburtenrate (0,02), während das Merkmal Milchproteinanteil (0,55) den höchsten Wert erreicht (FÜRST, 2021). Der Gesamtzuchtwert (RZG oder Relativer Zuchtwert Gesamt) setzt sich aus verschiedenen einzelnen Merkmalskomplexen zusammen. Diese Merkmalskomplexe stellen jeweils einen eigenen relativen Zuchtwert dar und werden je nach Zuchtziel unterschiedlich stark gewichtet. In Tabelle 1 sind die für die Zuchtwertschätzung relevanten Merkmale am Beispiel der Zuchtziele für das Deutsche Holstein dargestellt. Die Ergebnisse der Zuchtwertschätzung werden für Milchrinderrasen drei Mal pro Jahr (April, August und Dezember) und Fleischrinderrassen einmal pro Jahr (Dezember) veröffentlicht. Zusätzlich wurde 2014 der RZRobot, als Wert für die Melkbarkeit mittels automatisierter Melktechnik in die ZWS mit aufgenommen (VIT, 2021). Dieser wird u.a. durch Strichlänge und Strichplatzierung, Melkbarkeit, Zellzahl und Mobilität der Kühe bestimmt. Des Weiteren wurde die ZWS im April 2019 mit dem relativen Zuchtwert Gesund (RZGesund), definiert durch die Werte RZEuterfit (40 %), RZKlaue (20 %), RZRepro (15 %), RZMetabol (25 %) erweitert. Der RZKälberfit wurde ebenfalls im Jahr 2019 in die ZWS aufgenommen und spiegelt die Überlebenswahrscheinlichkeit weiblicher Kälber wider. Der Beobachtungszeitraum des Merkmals wird hierbei in fünf Abschnitte gegliedert und reicht insgesamt vom 3. Lebenstag bis zum 15. Lebensmonat (Zeitraum der ersten Besamung) eines jeden weiblichen Kalbes (siehe Tabelle 1). Zu beachten ist, dass die Überlebenschance eines Kalbes in den ersten 48 Stunden nach Geburt unabhängig vom Geschlecht im Merkmal Totgeburt aufgegriffen wird (VIT, 2021).

Tabelle 1: Relevante Merkmale der Zuchtwertschätzung beim Deutschen Holstein
Um die Zuchtziele der Rasse Deutsche Holsteins (DH) zu erreichen bzw. zu optimieren, wird der relative Gesamtzuchtwert (RZG) mit Hilfe der relativen Zuchtwerte der folgenden Merkmalskomplexe geschätzt. (Quelle: Genetische Vorgaben und relative Gewichtung der Merkmalskomplexe im RZG (VIT, 2021)

Merkmalskomplex [1]	%-Anteil des Merkmalskomplex im RZG für DH	Proportionale Gewichtung der Einzelmerkmale im Merkmalskomplex
Milchleistung (RZM)	36	Protein in kg (67 %) + Fett in kg (33 %)
Nutzungsdauer (RZN)	18	Funktionale Nutzungsdauer
Exterieur (RZE)	15	Milchtyp (10%) + Körper (20%) + Fundament (30%) + Euter (40%)
Fruchtbarkeit (RZR) [2]	7	Rastzeit Kühe (10%) + Non-Return-Rate-56 Kühe/Kalbinnen (37,5% / 7,5%)) + Verzögerungszeit Kühe/Kalbinnen (37,5% /7,5%) [3]
Maternale Kalbemerkmale (RZKm)	1,5	Maternaler Kalbeverlauf (33 %) + Totgeburt (67 %)
Paternale Kalbemerkmale (RZKp)	1,5	Paternaler Kalbeverlauf (33 %) + Totgeburt (67 %)
Gesundheit (RZG)	18	RZEuterfit (40 %) + RZKlaue (20 %) + RZRepro (15 %) + RZMetabol (25 %)
Kälberfitness (RZKälberfit)	3	Abschnitt 1 – 5 (je 20%) [4]

[1] RZ = relativer Zuchtwert (z.B.: RZM = relativer Zuchtwert Milchleistung)
[2] RZR = relativer Zuchtwert Reproduktion
[3] Rastzeit = Zeit von der Kalbung bis zur ersten Besamung; Non-Return-Rate-56 = 2. Besamung innerhalb von 56 Tagen (2 Zyklen) notwendig; Verzögerungszeit = Zeitraum zwischen erster Besamung und erstem Trächtigkeitstag
[4] Abschnitt 1 = 3. – 14. Lebenstag (Lt), Abschnitt 2 = 15. – 60. Lt, Abschnitt 3 = 61. – 120. Lt, Abschnitt 4 = 121. – 200. Lt, Abschnitt 5 = 201. - 458. Lt

2.1.3. Zuchtwertschätzung Töchterfruchtbarkeit

Zur Beurteilung der Fruchtbarkeit in der nachfolgenden Töchtergeneration wird der relative Zuchtwert Reproduktion (RZR) eines Bullen geschätzt. Zu diesem Zweck wird die Eigen- und Nachkommenleistung der vier Konzeptionsmerkmale (90 % des RZR) und der Rastzeit (10 % des RZR) beurteilt. Die Konzeptionsmerkmale werden getrennt für Kalbinnen und Kühe betrachtet und beinhalten die Non-Return-Rate 56 (NR56) und die Verzögerungszeit (VZ). Die Non-Return-Rate 56 gibt an, ob innerhalb von 56 Tagen eine erneute Besamung notwendig war, um eine Trächtigkeit herbeizuführen. Diese wird bei Kalbinnen und Kühen mit „ja" oder „nein" und bei Bullen mittels prozentualem Erstbesamungserfolg gewertet. Die prozentuale Gewichtung der Konzeptionsmerkmale sieht wie folgt aus: NR56 Kalbinnen 7,5 %, NR56 Kühe 36,5 % und VZ Kalbinnen 7,5 %, VZ Kühe 36,5 %. Für die Datenerhebung werden alle Belegungen ab 01.01.2000 sowie bei Kühen die sich wiederholenden Beobachtungen bis zur dritten Laktation berücksichtigt. Anschließend wird mittels BLUP (*best linear unbiased prediction*) Mehrmerkmals-Tiermodell der Zuchtwert geschätzt. Zu Korrektur werden verschiedenste fixe Umwelteffekte herangezogen, dazu zählen unter anderem Herde, Jahr, Belegungsmonat, Belegungsalter und verwendete Spermaart (Frischsamen oder Tiefgefriersperma). Der relative Zuchtwert wird mit einem Mittel von 100 und einer genetischen Standardabweichung von zwölf Punkten angegeben. Werte über 100 bedeuten für die Verzögerungszeit bzw. Rastzeit eine Verkürzung der Zeitspanne und wirken sich somit züchterisch positiv aus (VIT, 2021).

2.1.4. Zuchtwertschätzung Kalbemerkmale

Die für die ZWS der Kalbemerkmale verwendeten Eigen- und Nachkommenleistungen umfassen Informationen zum Kalbeverlauf und Totgeburten aus jeweils erster, zweiter und dritter Kalbung. Außerdem wird zwischen den paternalen und maternalen Einflüssen unterschieden. Paternale Effekte spiegeln die direkten Effekte eines Bullen auf das Kalb wider. Hierzu zählen unter anderem die Form und Größe eines Kalbes zum Zeitpunkt der Geburt. Die maternalen Effekte hingegen berücksichtigen den Einfluss eines Bullen auf die Kalbeeigenschaften seiner Töchter wie Größe, Neigungswinkel

und Durchmesser des Beckens. Der Verlauf der Kalbung selbst wird anhand vier Kalbeverlaufsklassen genauer spezifiziert. Diese reichen von leicht, normal, schwer bis zu nur unter tierärztlicher Hilfe bzw. Operation möglich. (RICHTER & GÖTZE, 1993; VIT, 2021).

Das Merkmal Totgeburt wird mit „ja“ oder „nein“ registriert und bezieht sich auf totgeborene oder innerhalb von 48 Stunden nach Geburt verendete Kälber. Zur Vereinfachung werden die Leistungsdaten des Kalbeverlaufs und der Totgeburt in einem Verhältnis von 1:2 jeweils für den paternalen und den maternalen Zuchtwert ermittelt (Tabelle 1). Die Schätzung des paternalen bzw. maternalen relativen Zuchtwerts erfolgt ebenfalls auf ein Mittel von 100 mit einer genetischen Standardabweichung von zwölf Punkten, wobei ein Wert über 100 weniger Schwer- und Totgeburten bedeutet und damit positiv gewertet wird. Die Schätzung wird ebenfalls mittels linearem Mehrmerkmalsmodell-BLUP-Tiermodell durchgeführt und enthält diverse fixe Effekte, wie Herde, Jahr und Kälbergeschlecht (VIT, 2021)

2.2. Ökonomische Kennzahlen

Das Holstein-Friesian Rind ist eine der ertragreichsten Milchviehrassen weltweit (RODRÍGUEZ-BERMÚDEZ et al., 2019). In Deutschland konnte 2019 ein durchschnittliche Milchleistung von 9.630 kg pro Schwarzbunter Holstein Kuh verzeichnet werden (VEREINIGTE INFORMATIONSSYSTEME TIERHALTUNG W.V., 2019). Das bedeutet, bei einem durchschnittlichen bundesweiten netto Milchpreis ab Hof von 33,70 Cent/kg (2019) konventionell erzeugter Milch konnte ein Landwirt durch den Milchverkauf durchschnittlich 3.245 € pro HF Kuh und Jahr erwirtschaften (VERBAND DER MILCHERZEUGER BAYERN E.V., 2019). Im Vergleich dazu erreichte das Fleckvieh im Jahr 2019 eine durchschnittliche Milchleistung von 7.246 kg Milch pro Kuh (VEREINIGTE INFORMATIONSSYSTEME TIERHALTUNG W.V., 2019). Verrechnet mit dem Milchpreis konnten somit mit einer Fleckvieh Kuh durchschnittlich nur 2.391 € durch den Milchverkauf eingenommen werden (VERBAND DER MILCHERZEUGER BAYERN E.V., 2019).

Neben den Daten der Milchleistungsprüfung wurden vom VIT Verden auch die Hauptabgangsursachen der Betriebe veröffentlicht. Hier geht klar hervor, dass Unfruchtbarkeit (18,5 %), gefolgt von Sonstigen Gründen (17,0 %), Verkauf zur Zucht (14,9 %) Klauen- bzw. Gliedmaßenerkrankungen (13,9 %) und Euterkrankheiten (13,0 %), die Hauptabgangsgründe im Jahr 2020 in den deutschen Milchviehbetrieben unabhängig von der gehaltenen Rasse waren (VIT, 2020). Die unzureichende Fruchtbarkeit äußert sich unter anderem in undeutlichen bis hin zu keinen Brunstsymptomen (schwache oder stille Brunst), Verlängerung verschiedener Fruchtbarkeitsintervalle (z.B. Güstzeit) oder einer schlechteren Konzeptionsrate (GRUNERT & BECHTOLD, 1999). Hinzu kommt der starke Einfluss auf das Tierwohl, der aus ethischer bzw. tierschutzrechtlicher Sicht ein wesentliches Problem darstellt. HUXLEY and WHAY (2006) konnten in ihrer Studie nachweisen, dass Schwer- und Totgeburten zu den traumatischsten Erlebnissen im Leben einer Kuh sowie des Kalbes zählen. Die Autoren beschrieben anhand einer Skala von 1 bis 10, wobei 10 den größten Schmerz darstellte, die Schmerzen einer Kuh während einer Schwergeburt mit 7. Im Falle des Kalbes wurde ein Wert von 4 erreicht. Diese Erlebnisse beeinflussen auch nachhaltig die Laktation des Muttertiers und die Entwicklung des Kalbes.

Darüber hinaus entstehen dem Landwirt beim Auftreten einer Schwer- und/oder Totgeburt hohe wirtschaftliche Einbußen. Die Kosten einer Geburt sind enorm vom Verlauf der Kalbung (KV) abhängig. Jede Kalbung kann einer der vier Kalbeverlaufs-Klassen zugeordnet werden: (1) leichte, (2) mittlere und (3) schwere Kalbung, sowie (4) Kalbungen, die tierärztliche Hilfe bzw. eine OP benötigen. Die Kosten pro Geburt setzen sich aus der zusätzlichen Arbeitszeit, Tierarzt- und Medikamentenkosten sowie den finanziellen Verlusten aufgrund verworfener Milch zusammen. In der nachstehenden Tabelle 2 sind die anfallenden Kosten je nach Kalbeverlaufsklasse im Detail aufgeführt. Hierbei wird deutlich, dass abhängig vom Schwierigkeitsgrad einer Kalbung der finanzielle Verlust eines Betriebes signifikant ansteigt (VEREINIGTE INFORMATIONSSYSTEME TIERHALTUNG W.V., 2020).

Im Falle einer Totgeburt (TG) bezieht sich der finanzielle Schaden des Landwirts im Wesentlichen auf den Verlust des Kalbes, das durchschnittlich einem Wert von 138 € entspricht. Je nach Verlauf der vorangegangenen Kalbung muss dieser Wert zu den Gesamtkosten des Kalbeverlaufs (Tabelle 2) addiert werden (VEREINIGTE INFORMATIONSSYSTEME TIERHALTUNG W.V., 2020).

Tabelle 2: Kostenaufstellung nach Kalbeverlauf
Die Tabelle stellt die anfallenden Kosten einer Geburt je nach Kalbeverlauf (KV-Klassen 1-4) dar. Hierbei handelt es sich um Durchschnittswerte, welche vom Vereinigte Informationssysteme Tierhaltung w.V. zur Verfügung gestellt wurden.

KV-Klasse [1]	Zusätzliche Arbeitszeit in Stunden	Zusätzliche Arbeitskosten [2]	Tierarztkosten	Medikamenten-kosten	Verworfene. Milch in kg [4]	Kosten der verworfenen Milch [6]	Gesamtkosten/ Fall
1	0	0	0	0	0	0	**0**
2	0,5	10	0	0	0	0	**10**
3	1,5	30	14 [3]	5	0	0	**49**
4	3	60	150	35	36 [5]	11,52	**256,52**

[1] Kalbeverlaufs-Klassen: 1 = leichte Kalbung; 2 = mittlere Kalbung, 3 = schwere Kalbung und 4 = Kalbung nur mit tierärztlicher Hilfe bzw. OP möglich
[2] Bei 20 €/h
[3] Annahme: bei 20 % der Schwergeburten wird ein Tierarzt hinzugezogen (à 70 €)
[4] Verworfene Milch gilt erst ab drei Tagen nach dem Kolostrum
[5] Annahme: bei einer Tagesleistung von 30 kg wird an 1,2 Tagen die Milch verworfen
[6] Milchpreis: 32 Cent/kg

Neben diesen fixen Kosten einer Geburt kann außerdem eine Aussage darüber getroffen werden, welche finanziellen Auswirkungen der Einsatz von Bullen mit negativen Zuchtwerten hat. Diese wirken sich entweder auf die direkten Effekte des Kalbes (paternaler Kalbeverlauf und Totgeburt) oder auf die Kalbeeigenschaften des Muttertiers (maternaler Kalbeverlauf und Totgeburt) aus. Die Kosten sind hierbei abhängig von der genetischen Standardabweichung des Zuchtwerts vom Mittel. Wie in Kapitel 2.1.4 beschrieben, wird der relative Zuchtwert der Merkmale Kalbeverlauf und Totgeburt mit einem Mittel von 100 und einer genetischen Standardabweichung von 12 Punkten angegeben (VIT, 2021). Die Kosten pro genetische Standardabweichung (gSD) betragen für das Merkmal maternaler/paternaler Kalbeverlauf 4,03 € bzw. 5,03 €, sowie für maternale/paternale Totgeburt 12,81 € bzw. 9,87 €. Diese Werte beziehen sich alle auf das Leben einer Kuh und entsprechen drei Laktationen (VEREINIGTE INFORMATIONSSYSTEME TIERHALTUNG W.V., 2020). Wird zum Beispiel ein Bulle mit einem Zuchtwert mTG = 72,51 ausgewählt, entspricht dies -2,29 genetische Standardabweichungen. Wird dieser Wert nun mit den Kosten pro 1 genetischen Standardabweichung multipliziert, erhält man einen Schaden von 29,33 € (d.h. 2,29 gSD × 12,81 €) pro Kuh, die mit diesem Bullen belegt wird. Zum Vergleich führt ein Bulle mit einem Zuchtwert mTG = 96,28 und einer gSD = 0,31 zu einem Verlust von 3,97 € pro Kuh.

2.3. Merkmale, molekulargenetische Marker, Genkarten und Kartierungsmethoden

2.3.1. Genetik quantitativer Merkmale

Entgegen der Mendelschen Regeln, nach welchen ein Gen nur ein Merkmal beeinflusst (z.B. Blütenfarbe), sind komplexe Merkmale (z.B. Krankheiten wie Diabetes beim Menschen, Milchleistung von Kühen, Krallenlänge von Vögeln) von mehreren genetischen und nicht-genetischen Mechanismen abhängig (GODDARD & HAYES, 2009). Die phänotypische Ausprägung von quantitativen Merkmalen wird sowohl auf genetischer als auch nicht-genetischer Ebene beeinflusst. Zu den genetischen Faktoren zählen u.a. das Vorkommen von Multiplen Allelen, Polygenie und Pleiotropie sowie der Wirkung gekoppelter Gene. Bei den nicht-genetischen Effekten spielen verschiedene Umwelteinflüsse eine Rolle bei der Ausprägung bestimmter Merkmale, z.B. Haltung und Fütterung (NICHOLAS, 2003). Merkmale können zudem an ein und demselben Genlocus unterschiedliche Allele aufweisen. Diese können funktionsunfähig (amorphe Allele), partiell funktionsfähig (hypomorphe Allele), gesteigert funktionsfähig (hypermorphe Allele), antagonistisch zum Wildtyp (antimorphe Allele) oder mit veränderten Eigenschaften (neomorphe Allele) in Erscheinung treten. Zusätzlich können diese unterschiedlichen Allele eines Gens unterschiedlich stark im Phänotyp erkennbar sein (GRAW, 2006). Die phänotypische Ausprägung der meisten Merkmale wird nicht durch ein Gen allein bestimmt, sondern von einer Vielzahl verschiedener Gene. Dieser Effekt ist als Polygenie oder multifaktorielle Vererbung bekannt (GRAW, 2006). Dementgegen steht die Pleiotropie, hier werden durch ein einzelnes Gen (genetische Ebene) mehrere Merkmale (phänotypische Ebene) beeinflusst (PAABY & ROCKMAN, 2013). Der Begriff der „antagonistischen Pleiotropie“ wurde erstmal von George C. Williams in seiner Forschung zum Alterungsprozess in den späten 50er Jahren erwähnt (WILLIAMS, 1957). Hierbei sind einzelne Gene in der Lage, sich gegenseitig zu beeinflussen. Beispielsweise kann Gen X negativen Einfluss auf die Genexpression des Gens Y und positive Effekte auf die des Gens Z nehmen (LEROI et al., 2005).

2.3.2. Molekulargenetische Marker

Genetische Marker stellen Bezugspunkte im Genom dar und werden zur Erstellung physikalischer und genetischer Genkarten als auch zur Kartierung kausaler Loci monogener und polygener Merkmale verwendet (GRAW, 2006). Bei diesen genetischen Markern handelt es sich um variable, locusspezifische DNA-Sequenzen, welche aus einer variablen und einer nicht-variablen Komponente bestehen. Der nicht-variable Teil des Markers dient dazu, den Marker an einem einzelnen Locus zu identifizieren, während der variable Teil Unterschiede zwischen Individuen als auch zwischen den homologen Chromosomen eines Tieres widerspiegelt (BURTON et al., 2005). Des Weiteren können DNA-Marker in zwei Typen eingeteilt werden. Einerseits Typ-I-Loci Marker, welche sich innerhalb oder eng flankierend an der relevanten Genregion befinden und andererseits die der Typ-II-Loci mit unbekannter Funktion (GELDERMANN, 2005). Aufgrund der Unabhängigkeit genetischer Marker von Geschlecht, Alter, Gewebe und verschiedenen Umwelteinflüssen eignen sich diese hervorragend für verschiedenste genetische Fragestellungen und Kartierungsansätze (GRAW, 2006).

Im Rahmen der genetischen Kartierung hängt der Nutzen genetischer Marker vom Grad ihrer Polymorphie an den einzelnen Loci ab. Das bedeutet, nur, wenn an einem Locus mindestens zwei unterschiedliche Allele sowie heterozygote Genotypen vorliegen, können darüber Rückschlüsse auf Rekombinationsereignisse zwischen Generationen gezogen werden. Des Weiteren steigt der informative Nutzen eines Markers, wenn die verschiedenen Allele in einer Population annähernd gleich oft auftreten (d.h. ähnliche Allelfrequenz) (GELDERMANN, 2005). Darüber hinaus kann umso genauer kartiert werden, desto mehr Marker ein Chromosom bzw. das Genom abdecken (GEORGES, 2007). Typischerweise wurden lange Zeit Mikrosatellitenmarker verwendet, allerdings konnten die kartierten Loci nur auf einen relativ großen Bereich von mehr als 20 cM (Centimorgan) eingegrenzt werden. Mit diesem großen Fenster war es nur schwer möglich, kausale Gene monogener und polygener Merkmale zu detektieren (GODDARD & HAYES, 2009). Eine bessere Eingrenzung der Kandidatenloci ist mit der Verwendung von SNPs (*single nucleotide polymorphisms*) als Marker möglich, wodurch diese heutzutage hauptsächlich in Kartierungsanalysen verwendet werden (WELLER, 2001). Aus diesem Grund wird im Folgenden

nur auf die Einzelnukleotid-Polymorphismen näher eingegangen.

2.3.2.1. Single Nucleotide Polymorphisms (SNPs)

SNPs sind einfachste Genvarianten (Polymorphismen), die durch Basen-Substitutionen entstehen (GRAW, 2006). Bei dieser Form der Mutation wird durch ein zufälliges Mutationsereignis eine ursprüngliche Base (Wildtyp- oder *ancestral*-Allele) durch eine neue Base (mutierte oder *derived*-Allele) ersetzt. Wird eine Pyrimidinbase (Thymin, Cytosin und Uracil) durch eine andere Pyrimidinbase oder eine Purinbase (Adenin und Guanin) durch eine andere Purinbase ausgetauscht, handelt es sich um eine Transition. Bei der Transversion erfolgt der Austausch einer Pyrimidinbase durch eine Purinbase und umgekehrt (VIGNAL et al., 2002). Je nach Spezies und Chromosomenposition erscheinen Einzelnukleotid-Polymorphismen in Abständen von 0,1–1 kb entlang des DNA-Stranges und treten meist biallelisch auf (GELDERMANN, 2005). Für gewöhnlich befinden sich diese Varianten in Bereichen zwischen Genen und haben nur einen sehr geringen bis gar keinen Effekt. Andere Varianten befinden sich innerhalb kodierender Regionen und können entweder als stille, Missense- oder Nonsense-Mutation in Erscheinung treten (NICHOLAS, 2003). Während es bei stillen Mutationen zu keinen Veränderungen der abgeleiteten Aminosäure kommt, wird bei der Missense-Mutation die Aminosäure ausgetauscht und im Falle der Nonsense-Mutation ein Stopp-Codon erzeugt und in Folge der Abbruch der Proteinsynthese bewirkt (GRAW, 2006).

Trotz der geringeren Informativität der biallelischen SNP-Marker sind diese in der Gendiagnostik den multi-allelischen Mikrosatelliten-Markern überlegen (VIGNAL et al., 2002). Die genaue Kartierung kausaler Loci ist nur in Bereichen mit hoher Markerdichte möglich (GEORGES, 2007). Hier sind SNPs klar im Vorteil, da diese auch in Mikrosatelliten-freien Regionen vorkommen. Zusätzlich erwiesen sich SNPs nicht nur als stabiler, sondern auch weniger anfällig gegenüber Mutationen. Das bedeutet, ein SNP Allel „A“ bleibt selbst in großen Datensätzen bzw. über mehrere Generationen hinweg Allel „A“ (WELLER, 2001). Des Weiteren ist aufgrund der einfachen Bestimmung der SNP Allele, d.h. Allel entspricht der Referenz oder nicht, eine automatisierte Genotypisierung möglich, wodurch große Datensätze zeit- und kosteneffizient

genotypisiert werden können (VIGNAL et al., 2002). Die Firma Illumina Inc. stellt für das Rindergenom zwei verschiedene SNP-Chips zur Verfügung, den BovineSNP50 BeadChip und den BovineHD BeadChip. Die Marker-Chips unterscheiden sich sowohl in der Anzahl der enthaltenen SNPs als auch der Menge an Proben, die gleichzeitig genotypisiert werden können. Während mit Hilfe des 50K BeadChips der Version 3 zeitgleich in 24 Proben jeweils 53.218 SNPs genotypisiert werden (ILLUMINA INC., 2020a), sind es bei Verwendung des HD BeadChips 777.962 SNPs in jeweils 8 Proben. Dadurch deckt der HD BeadChip beinahe das gesamte Rindergenom ab (ILUMINA INC., 2015).

2.3.3. Genkarten

Die Genomkartierung kann mittels physikalischer oder genetischer Kartierung erfolgen. Der Unterschied beruht auf der Messung der absoluten Abstände von DNA-Markern in Basenpaaren (bp) auf der physikalischen Karte und der relativen Abstände in Centimorgan (cM, nach Thomas Hunt Morgan) bei einer genetischen Kartierung (GELDERMANN, 2005).

Bei der früheren physikalischen Kartierung wurde die absolute Position eines Markers auf dem DNA-Strang anhand der Abstände zwischen den Chromosomenbändern ermittelt und auf einer Karte dargestellt. Gängige zytogenetische Verfahren hierfür waren die lichtmikroskopisch gestützte Chromosomenbänderung, Zellhybridisierungstechniken und DNA-In-Situ-Hybridisierung (GELDERMANN, 2005). Die moderne und ultimative physikalische Kartierung misst die Anzahl an Nukleinbasen zwischen Genen oder Markern. Die Erstellung einer solchen Karte basiert auf der Ganzgenomsequenzierung (*Whole-genome sequencing – WGS*) eines Individuums. Mit Hilfe der derzeitigen Sequenzierungsmethoden ist es bisher jedoch nicht möglich, die Sequenz des gesamten Genoms an einem Stück zu sequenzieren. Daher werden zur Rekonstruktion der Genomsequenz die DNA-Fragmente (*Reads*) der Sequenzierung anhand überlappender Abschnitte wieder aneinandergereiht. Diese Technik ist als *de novo* Assemblierung bekannt (siehe Kapitel 3.2.5.4) (NG & KIRKNESS, 2010).

Im Falle der genetischen Kartierung wird mittels Kopplungsanalysen (*engl. Linkage Mapping*) die Rekombinationshäufigkeit zwischen zwei Markergenen

ermittelt. Hierbei gilt, umso weiter zwei Loci auf einem Chromosom voneinander entfernt liegen, desto häufiger treten Rekombinationsereignisse zwischen diesen auf (GRAW, 2006). Alfred Harry Sturtevant (1891-1970) arbeitete als Promotionsstudent im Fliegenzimmer von T. H. Morgan und erkannte 1911, dass sich anhand dieser Wahrscheinlichkeiten die relative Lage von Genen auf einem Chromosom bestimmen lässt und somit genetische Chromosomenkarten erstellt werden können. Hierfür wird die relative Anzahl an Rekombinationsereignissen innerhalb einer Meiose als Abstand zweier Merkmale gewertet und mit einem cM auf der Kopplungskarte eingetragen. Eine Centimorgan-Einheit entspricht 1 % Rekombination, d.h. bei jeder hundertsten Meiose tritt eine Rekombination auf (GRAW, 2006). Wenn für einen physikalisch kartierten Markerlocus verschiedene Allele bekannt sind, ist ein Vergleich der beiden Kartierungsmethoden möglich, hierbei entspricht 1 cM bei vielen Säugetierspezies im Durchschnitt etwa 10^3 kb (GELDERMANN, 2005).

2.3.4. Kopplung und Rekombination von Genen

Thomas Hunt Morgan erkannte in einem Kreuzungsexperiment, dass Gene bestimmter Merkmale in den Nachkommen stets zusammenblieben und widerlegte somit die dritte Mendelsche Regel (Unabhängigkeitsregel). Diese besagt, dass sich Allele eines Gens sowohl unabhängig voneinander als auch unabhängig von Allelen anderer Gene in der nachfolgenden Generation verteilen. Die von T. H. Morgan beschriebene gemeinsame Vererbung von Genen wird als Kopplung (*linkage*) bezeichnet und tritt umso häufiger auf, je näher die Gene eines Merkmals auf einem Chromosom beieinanderliegen. Eine Neuverteilung der gekoppelten Gene ist wiederum nur möglich, wenn durch Rekombinationsereignisse (*Crossing-over*) zwischen homologen Chromosomen ein Austausch der gekoppelten Allele erfolgt (GRAW, 2006). Können dieselben Kopplungsgruppen an denselben Loci verschiedener Individuen einer Population oder in anderen Rassen bzw. Spezies nachgewiesen werden, können hierüber Rückschlüsse auf deren genetische Verwandtschaft gezogen werden (GELDERMANN, 2005).

Werden zwei Loci nur zufällig miteinander vererbt, so gilt, dass die Allelfrequenz der gekoppelten Loci gleich der Allelfrequenz der einzelnen Loci in einer Population ist. Mit anderen Worten: die Allele der Loci treten sowohl gekoppelt als auch einzeln gleich oft in der Population auf und befinden sich daher in einem Kopplungsgleichgewicht (*linkage equilibrium*) (TERWILLIGER et al., 1998). Sind die Allele zweier Loci jedoch nicht zufällig miteinander gekoppelt, sind diese im Kopplungsungleichgewicht (*linkage disequilibrium, LD*) zueinander. Die Allelfrequenz der gekoppelten Loci ist in daher entweder höher oder niedriger als die Allelfrequenz der einzelnen Allele in einer Population (PRITCHARD & PRZEWORSKI, 2001). Die nicht zufällige Kopplung basiert auf vergangenen Evolutionsfaktoren in einer Population wie Mutationen, Selektion, Gendrift- und Genshift- Ereignissen sowie assortativer Anpaarung (WU & ZENG, 2001). Durch Rekombinationen nimmt das LD der Allele mit der Zeit jedoch wieder ab. Auch hier gilt wieder, umso näher sich die Allele auf dem Chromosom zueinander befinden, desto seltener kommt es zur Rekombination und das Kopplungsungleichgewicht bleibt länger bestehen (OLSEN et al., 2004). Daher indiziert ein starkes Kopplungsungleichgewicht auch eine physisch nahe Kopplung der Loci (KAPLAN et al., 1995).

2.3.4.1. Kopplungsanalyse

Die Intention der Kopplungsanalyse ist es also, zu ermitteln, ob zwei Loci eines Chromosoms abhängig voneinander vererbt werden. Zu beachten ist, dass bei Individuen, die an Loci homozygote Allele aufweisen, keine Aussage über eine mögliche Rekombination getroffen werden kann, weshalb diese für die betreffende Kopplungsanalyse nicht informativ sind. Die Rekombinationshäufigkeit zweier Loci wird mittels der Rekombinationsfraktion θ angegeben und im Zuge der Kopplungsanalyse geschätzt. Bei ungekoppelten Loci ist $\theta = 0{,}5$, bei gekoppelte Loci gilt $\theta < 0{,}5$.

Die Kopplungsanalyse wird unter anderem zur Kartierung der kausalen Loci monogener oder polygener Merkmale angewendet. Im Falle polygener Merkmale handelt es sich um die sogenannte QTL-Kartierung. Hierbei wird eine große Anzahl an Markern genutzt, deren Positionen auf dem Genom

bekannt sind, d.h. eine Markerkarte existiert (TERWILLIGER & GORING, 2000). Zur Kartierung wird die Kopplung der flankierenden Marker zu einem hypothetischen QTL geschätzt. Hierbei gilt, umso enger der Marker mit dem QTL gekoppelt ist, desto wahrscheinlicher treten die gleichen Marker-QTL Kopplungsphasen (Haplotypen) auch in den nachfolgenden Generationen auf (GELDERMANN, 2005). Jedoch hat die Kopplungsanalyse zwei wesentliche Nachteile. Zum einen liefern nur die Proben verwandter Individuen zuverlässige Ergebnisse. Zum anderen sind stark gekoppelte Marker seltener von Rekombinationsereignissen betroffen, wodurch kausale Loci nur auf einen großen Bereich von mehreren Megabasen eingegrenzt werden können (COLLINS, 2009). Um einen Loci nun präziser kartieren zu können, wird auf eine Feinkartierung mittels LD-Analyse zurückgegriffen (MEUWISSEN & GODDARD, 2000).

2.3.4.2. Kopplungsungleichgewichtsanalyse

Ähnlich der Kopplungsanalyse ist es das Ziel der Kopplungsungleichgewichtanalyse, auch bekannt als Assoziationskartierung (RISCH & MERIKANGAS, 1996), kausale Loci mit Effekten auf ein bestimmtes Merkmal zu kartieren. Jedoch wird die LD-Analyse mit historischen Pedigree-Informationen aus ganzen Populationen mehrerer Generationen durchgeführt, während die Kopplungsanalyse auf Pedigree-Daten genotypisierter Tiere zurückgreift (TERWILLIGER & GORING, 2000). Durch die Nutzung zahlreicher historischer Rekombinationsereignisse und die einfachere Probengewinnung (d.h. Unabhängigkeit von Stammbauminformationen) ist mit der LD-Analyse eine exaktere Kartierung von Kandidatenloci im Gegensatz zur Kopplungsanalyse möglich (COLLINS, 2009).

Die Kartierung der Kopplungsungleichgewichtsanalyse basiert auf der Identifikation chromosomaler Regionen, die aufgrund ihrer gemeinsamen Herkunft (*identity by descent, IBD*) identisch sind (MEUWISSEN & GODDARD, 2000). Hierbei werden vornehmlich solche Tiere zur Untersuchung ausgewählt, die den Phänotyp des Merkmals deutlich widerspiegeln (TERWILLIGER & GORING, 2000). Als Bezugspunkte im

Genom dienen wie schon in der Kopplungsanalyse genetische Marker (KRUGLYAK, 1999). Bei Vorliegen zweier identischer Marker-Haplotypen in einem oder mehreren Merkmalsträgern besteht eine höhere Wahrscheinlichkeit, dass diese chromosomalen Abschnitte herkunftsidentisch sind. Dementsprechend sind auch die unbeobachteten Loci zwischen den Markern IBD und stehen im Kopplungsungleichgewicht zu dem Merkmal (MEUWISSEN et al., 2002). Assoziationen zwischen Markern und Merkmalen sollten jedoch innerhalb einer gesamten Population existieren und idealerweise nicht verwandtschaftsabhängig auftreten. Dies ist schwer zu überprüfen, da die meisten Nutztierpopulationen auf Halb- (Rind) bzw. Vollgeschwister-Strukturen (Schwein) basieren. Folglich treten Kopplungsungleichgewichte zwischen Loci auf, die im Prinzip nicht gekoppelt wären. Dies lässt sich folgendermaßen erklären: Wenn ein Vatertier seltene Allele an zwei ungekoppelten Regionen aufweist, werden diese beiden Allele bei seinen Nachkommen mit höherer Wahrscheinlichkeit auftreten, als dies bei unverwandten Tieren der Fall wäre. Wird eines der beiden seltenen Allele als QTL detektiert, erscheint er in LD mit dem zweiten Locus. Um derartig falsch positive Ergebnisse zu vermeiden, wird eine Variable in das Modell eingefügt, die all jene Gene berücksichtigt, die einen Einfluss auf das Merkmal haben (polygenetischer Effekt) (GODDARD & HAYES, 2009).

2.3.4.3. Kombinierte Kopplungsungleichgewichts- und Kopplungsanalyse (*Combined Linkage Disequilibrium and Linkage Analysis; cLDLA*)

In den vorangegangenen Abschnitten wurden zwei Methoden zur Kartierung quantitativer Merkmale kurz vorgestellt. Um die Informationen beider Methoden kombiniert miteinander nutzen zu können, beschrieben sowohl MEUWISSEN et al. (2002) als auch FARNIR et al. (2002) ein Verfahren, das gerade im Bereich der Feinkartierung von kausalen Genen mit Effekten auf quantitative Merkmale Anwendung findet. Die cLDLA beruht auf dem in Kapitel 2.3.4.2 beschriebenen Prinzip der IBD-Wahrscheinlichkeiten zwischen Marker-Haplotypen und den unbeobachteten QTL Allelen mit Effekten auf das zu untersuchende quantitative Merkmal (FARNIR et al., 2002). Innerhalb des bekannten Pedigrees kann die IBD-Wahrscheinlichkeit

mittels Kopplungsanalyse ermittelt werden. Sollte es sich allerdings um unverwandte Tiere bzw. Tiere mit unbekanntem Pedigree handeln, wird die IBD-Wahrscheinlichkeit mittels Kopplungsungleichgewichtsanalyse geschätzt (MEUWISSEN et al., 2002). Die bereits angesprochenen falsch positiven Ergebnisse der Kopplungsungleichgewichtsanalyse können durch Haplotypen-basierte Verfahren bei gleichzeitiger Steigerung der Kartierungsgenauigkeit minimiert werden (FARNIR et al., 2002).

Die in dieser Studie angewendete Methode basiert vor allem auf der Arbeit von MEUWISSEN et al. (2002) und wird im Kapitel 3.2.3 genauer beschrieben.

2.4. Sequenzierungsmethoden der dritten Generation

Die Sequenzierung nach Sanger war eine der ersten Methoden zur Entschlüsselung ganzer Genomen und folglich der Detektion von Genen bzw. Mutationen (POLLARD et al., 2018). Fragmente mit einer durchschnittlichen Länge von 600 Basenpaaren (AMARASINGHE et al., 2020) wurden bei der Methode nach Sanger innerhalb eines Durchganges einzeln sequenziert (ILLUMINA INC., 2020b). Jedoch eignete sich diese Methode aufgrund des enormen Zeit- und Arbeitsaufwandes lediglich zur Darstellung einzelner kurzer Chromosomenabschnitte. Als nächster großer Entwicklungsschritt folgte die Next-Generation Sequenzierung (NGS). Mit dieser Methode war es nun möglich, Millionen von Reads gleichzeitig zu sequenzieren. Zusätzlich konnte die Genauigkeit durch die Paired-end-Sequenzierung weiter gesteigert werden. Hierbei werden die Reads beider DNA-Orientierungen (Positiv- und Minusstrang) gleichermaßen gegen das Referenzgenom gemappt. In Folge stehen mehr und längere Reads für ein genaueres Mapping gegen das Referenzgenom zur Verfügung und einfache Strukturvarianten können im Gegensatz zur Nutzung von Single-Read Daten, welche nur eine DNA Orientierung beachten, dargestellt werden (ILLUMINA INC., 2015).

Einen weiteren Entwicklungsschritt stellte die Third-Generation bzw. Long-Read Sequenzierung dar. Gängige Unternehmen bzw. Verfahren sind Pacific Biosciences (PacBio), Single-molecule real-time sequencing (SMRT) und die Oxford Nanopore Technologie (ONT). Mit Hilfe der Long-Read Sequenzierung können Fragmente mit einer durchschnittlichen Basenlänge von 10.000 bp produziert werden (AMARASINGHE et al., 2020). Mit dieser Entwicklung wurde eine Vielzahl neuer Analyseansätze ermöglicht, die zuvor mittels NGS-Verfahren nicht in der gleichen Qualität durchführbar waren. Dazu zählen die Erstellung neuer Assemblies (*de novo* Assembling) oder die gezielte Sequenzierung komplexer Regionen, wie z.B. in Bereichen der HLA-Komplexe (Histocompatibility Leucocyte Antigen) oder von Genen wie dem *ADPKD* (*Autosomal-dominant polycystic kidney disease*). Des Weiteren ermöglichen lange DNA-Fragmente die Sequenzierung von Isoformen sowie die genauere Untersuchung von Splicing Vorgängen. Auch Strukturvarianten, wie z.B. Duplikationen, weitreichende Deletionen, die mit dem Verlust einzelner Gene einhergehen, oder Fusionen bestimmter Chromosomenbereiche, können mittels Third-Generation Sequenzierung

identifiziert werden. Außerdem ist eine Charakterisierung der DNA-Methylierung, klonaler Heterogenität verschiedener Pathogene und von Immunzellen möglich (POLLARD et al., 2018).

2.5. Strukturelle Chromosomenaberrationen

Strukturelle Chromosomenaberrationen können spontan oder induziert auftreten, z.B. durch Röntgenstrahlung oder chemische Mutagenese (ROTHWELL, 1993). Sie erscheinen unter anderem in Form von Deletionen, Duplikationen, Inversionen und Translokationen (GRAW, 2006). Einfache strukturelle Varianten (SV) entstehen durch ein, zwei oder mehr Doppelstrangbrüche innerhalb eines Chromosoms bzw. mit Transfer zu einem anderen Chromosom. Komplexe Aberrationen weisen hingegen deutlich mehr Brüche auf und stellen häufig eine Kombination mehrerer einfacher Strukturvarianten dar (WECKSELBLATT & RUDD, 2015). Um das Zentromer bzw. das Telomer und damit die Funktionalität der DNA zu erhalten, werden verschiedene Reparationsvorgänge dieser Brüche unternommen. In Folge fehlerhafter Reparationsversuche ist einerseits eine Umstrukturierung der Genanordnung und andererseits eine veränderte Basensequenz der Gene zu beobachten (GRAW, 2006).

2.5.1. Deletionen und Duplikationen

In Folge eines Doppelstrangbruchs (DSB) können innerhalb (interstitiell) (*abcdefg* → *abfg*) oder am Ende (terminal) (*abcdefg* → *abcd*) eines Chromosoms Bruchstücke verloren gehen (Deletion). Abhängig von der Größe des deletierten Fragments können diese unterschiedlich schwere Auswirkungen haben (ROTHWELL, 1993). Um die Funktionalität und Integrität der DNA zu erhalten, finden in eukaryotischen Zellen hauptsächlich zwei Reparationsvorgänge von Doppelstrangbrüchen statt. Einerseits wird mittels homologer Rekombination (HR) und andererseits durch nicht-homologe End-zu-End-Verknüpfung (non-homologous end joining – NHEJ) der Doppelstrangbruch der DNA behoben. Bei der homologen Rekombination wird die Bruchstelle unter Zuhilfenahme des homologen Stranges rekonstruiert. Im Falle der NHEJ können mit Hilfe der DNA-Ligase DNA-Enden wieder aneinandergereiht werden. Dies geschieht ohne jegliche Folgen für die DNA bei kohäsiven Endstücken. Jedoch können vereinzelt Insertionen und Deletionen durch die Reparationsvorgänge entstehen, welche Veränderungen der Basenabfolge einer Genomsequenz und folglich der

Proteinexpression nach sich ziehen (KOSZUL & FISCHER, 2009). Weitreichende, homozygote oder innerhalb kritischer Gene lokalisierte Deletionen enden meist letal für das Individuum (ROTHWELL, 1993). Des Weiteren kann eine sogenannte Pseudodominanz bestimmter Allele beobachtet werden. Hierbei tritt durch Deletion des dominanten Alleles und damit Verlust des dominanten Faktors, das rezessive Allel phänotypisch in Erscheinung (GRAW, 2006).

Eine weitere komplexere Mutationsform stellen segmentale Duplikationen (SD) oder Low Copy Repeats (LCR) dar, welche komplizierte Chromosomenaberrationen (z.B. Copy Number Variationen) nach sich ziehen können. SD zeichnen sich dadurch aus, dass die duplizierten Abschnitte zu mindestens 90 % deckungsgleich und eine Mindestlänge von 5 kb aufweisen (SHARP et al., 2005). Die im Genom nachgewiesenen SD werden nach ihrer Verteilung und Organisation in drei Klassen eingeteilt: perizentrische, um das Zentromer gelegen, subtelomerische, an den äußeren Enden eines Chromosoms oder interstitielle segmentale Duplikationen, die übrigen Bereiche innerhalb des Chromosoms betreffend. Zusätzlich können sich die duplizierten Segmente auf ein und demselben Chromatid, auf den beiden Schwesterchromatiden oder auf zwei Chromosomen befinden (BAILEY & EICHLER, 2006). Probleme treten dann zum Beispiel während dem Crossing-over auf, wenn sich eine nahezu identische Basensequenz auf dem gleichen Chromatid, dem Schwesterchromatid oder dem homologen Chromosom befindet (KOSZUL & FISCHER, 2009). Aufgrund der identischen Basensequenz werden duplizierte Abschnitte während dem Crossing-over durch z.B. Schleifen aneinandergelagert. Dieses Phänomen wird als nicht-allelische homologe Rekombination (NAHR) bezeichnet. Die NAHR kann intrachromatid (auf ein und demselben Chromatid), intrachromosomal (zwischen den beiden Schwesterchromatiden) und interchromosomal (zwischen zwei Chromosomen) stattfinden und zieht unter anderem Deletionen und Duplikationen auf einem oder beiden Chromatiden bzw. in homologen Chromosom nach sich. Darüber hinaus könne sowohl dizentrische als auch azentrische Chromatide entstehen (Abbildung 1) (STANKIEWICZ & LUPSKI, 2002). Dizentrische Chromosomen enthalten zwei Zentromere und bilden zwischen diesen eine sogenannte dizentrische Brücke aus, zudem treten bestimmte Regionen gegenüber anderen vermehrt auf. In den

azentrischen Chromosomen kann kein Zentromer nachgewiesen werden und es fehlen Teile der DNA-Sequenz (ROTHWELL, 1993). Da das azentrische Chromosom aufgrund des fehlenden Zentromers nicht von den Spindelfasern erfasst werden kann, geht dieses verloren und die Zygote ist somit nicht lebensfähig (JANNING & KNUST, 2004).

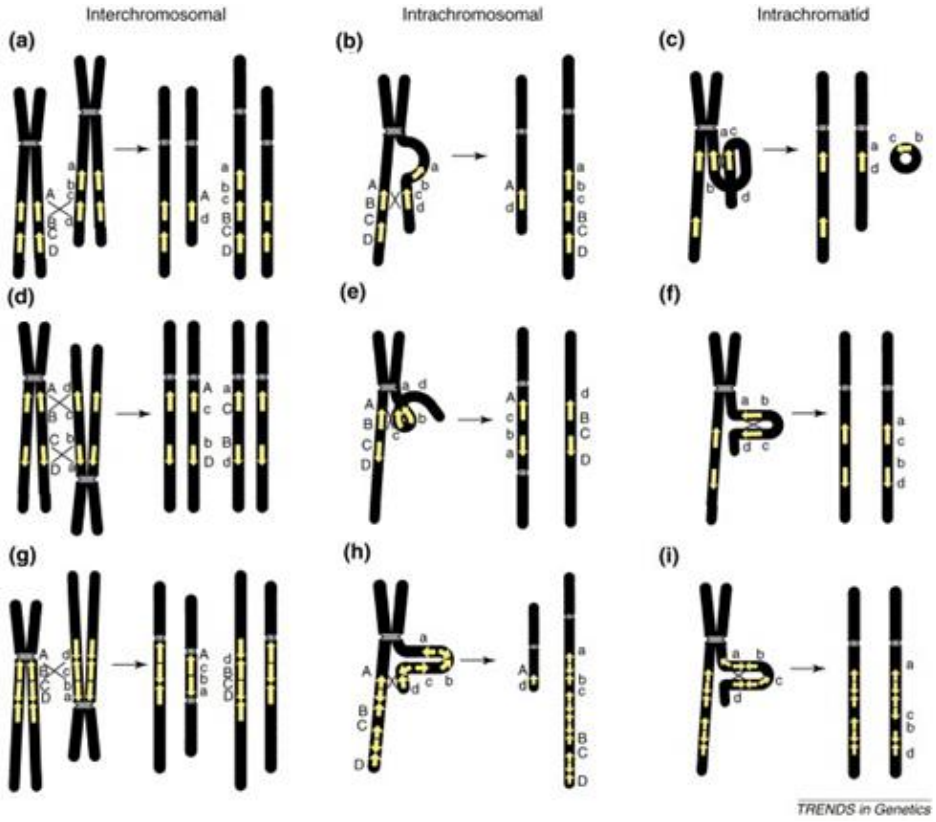


Abbildung 1: Auswirkungen der nicht-allelischen homologen Rekombination (NAHR) auf die genomische Struktur
(Abbildung modifiziert nach STANKIEWICZ and LUPSKI (2002))
Die Chromosomen sind in schwarz, die Zentromere grau und segmentale Duplikationen in gelb dargestellt. Gleichermaßen orientierte segmentale Duplikationen (SD) führen zu Deletionen bzw. Duplikationen in homologen Chromosomen (a), oder den beiden Schwesterchromatiden (b). Erfolgt die NAHR auf ein und demselben Chromatid, enthält ein Teil dieses Chromatids eine Deletion und es entsteht ein azentrisches Fragment (c). Die NAHR SD mit entgegengesetzter Orientierung äußert sich durch Inversionen (d-f) sowie der Ausbildung eines dizentrischen und eines azentrischen Chromatids (e), wohingegen komplexe SD eine komplizierte Kombination aus Inversionen, Deletionen und Duplikationen nach sich ziehen (g-i).

2.5.2. Inversionen

Bei der Inversion wird ein Teilstück aus einem Doppelstrangbruch um 180° gedreht wieder eingesetzt (*abcdefg* → *ab**fedcg***). Des Weiteren wird eine perizentrische, d.h. das Zentromer ist innerhalb der Inversion, von einer parazentrischen Inversion, d.h. das Zentromer liegt außerhalb der Inversionsstelle, unterschieden (JANNING & KNUST, 2004). Individuen mit

einer heterozygoten Inversion sind genetisch ausgeglichen und ohne jeglichen pathologischen Effekt, da die Gene nicht in ihrer Anzahl, sondern nur der Anordnung verändert sind (ROTHWELL, 1993). Allerdings können die Bruchstücke an den neuen Enden nicht mehr auf die gleiche Weise miteinander interagieren und eine veränderte Genwirkung tritt auf. Dies geschieht vor allem dann, wenn Gene aus aktiven euchromatischen Bereichen in inaktive heterochromatische Abschnitte verlagert werden (GRAW, 2006). Dieser sogenannte Positionseffekt kann sowohl bei Inversionen, als auch bei Translokationen, Deletionen und Duplikationen vorkommen (JANNING & KNUST, 2004). Ähnlich der Deletion und Duplikation zeigen sich die negativen Auswirkungen der Inversion erst bei der Paarung und dem Crossing-over. Um exakte homologe DNA-Stränge zu gewährleisten, werden Schleifen im Bereich der Inversion gebildet. Anschließend finden innerhalb dieser Schleifen das Crossing-over statt. Die zytologischen Auswirkungen sind abhängig davon, ob eine peri- oder parazentrische Inversion vorliegt (JANNING & KNUST, 2004). Bei einer heterozygoten perizentrischen Inversion liegt das Zentromer innerhalb der Bruchstellen. Es entstehen zwei morphologisch unauffällige Chromosomen, allerdings beinhalten diesen sowohl Duplikationen als auch Deletionen. Aus *ABCDEFG* und *abfedcg* wird einmal *ABCDEf**ba*** und *gcdeF**G***. Das erste Chromosom besitzt dabei zweimal den Abschnitt ***ab*** bei Fehlen von ***g*** und das Zweite unterliegt einer Duplikation von ***g*** und Deletion des Anteils ***AB***. Nur durch ein erneutes Crossing-over (d.h. Doppelcrossover) der gleichen Chromatiden kann das genetische Ungleichgewicht verhindert werden. Betroffene Individuen bzw. Populationen einer perizentrischen Inversionen zeigen partielle Fertilitätsverluste (ROTHWELL, 1993). Als Folge der heterozygoten parazentrischen Inversion entstehen nach der Rekombination dizentrische und azentrischen Chromosomen, die zu einer lebensunfähigen Zygote führen (siehe oben Kapitel 2.5.1).

2.5.3. Translokationen

Durch zwei oder mehr Brüche im selben oder verschiedenen Chromosomen können Segmente an neuen Positionen wieder eingebaut werden. Gelegentlich werden ganze Chromosomenarme an das Ende eines anderen

Chromosoms verlagert, hierbei handelt es sich um eine Tandem-Translokation (NICHOLAS, 2003). Die einfachste Form ist der genetic-shift in ein und demselben Chromosom. Nach einem dreimaligen Bruch wird das Bruchstück an einem anderen Locus wieder eingesetzt (*abcdefg* → *ab**ef**cdg*) (ROTHWELL, 1993). Ein Austausch zwischen zwei nicht homologen Chromosomen (z.B. BTA18 und 27) wird als reziproke Translokation bezeichnet, in diesem Fall enthalten nur beide Translokationschromosomen alle Gene. Individuen die Translokationen innerhalb eines Chromosoms aufweisen sind i.d.R. phänotypisch unauffällig und es treten keinerlei Probleme während der Paarung und der Rekombination auf, solange die Translokation im diploiden Satz homozygot vorliegt (JANNING & KNUST, 2004). Erst bei heterozygoten Translokationen sind in der Meiose Prophasenpaarungsfiguren notwendig (GRAW, 2006). Dieser Fall tritt ein, wenn genetisch ausbalancierte Gameten mit einer Translokation mit normalen Gameten kombiniert werden und die Nachkommen jetzt eine heterozygote Translokation aufweisen (NICHOLAS, 2003). Um eine Paarung der homologen Chromosomenabschnitte zu gewährleisten, wird eine kreuzförmige Doppeltetrade ausgebildet (JANNING & KNUST, 2004). Das Hauptproblem stellt hier nicht das Crossing-over dar, sondern die fehlerhafte Trennung in der anschließenden Anaphase der Meiose, abhängig von der räumlichen Nähe der Bruchstücke zum Zentromer (ROTHWELL, 1993). Die produzierten Gameten sind genetisch im Ungleichgewicht für den Translokationsbereich, welcher in den einen dupliziert und in den anderen defizient vorliegt (JANNING & KNUST, 2004). Bei einer Verschmelzung mit einem anderen Gameten zu einer Zygote ist diese nicht lebensfähig und der Embryo stirbt (NICHOLAS, 2003).

2.6. Epigenetik

Mehrere tausend Gene eines Organismus werden in nahezu allen Zellen auf die gleiche Weise transkribiert. Diese Gene werden als sogenannte Haushaltsgene (*engl. housekeeping genes*) bezeichnet. Daneben existieren Gene, die entweder nur in ausgewählten Zelltypen, in bestimmten Lebens- oder Zyklusphasen oder nur unter bestimmten Umwelteinflüssen aktiv sind (JANNING & KNUST, 2004). Die unterschiedliche Genexpression wird bereits während der embryonalen Zellentwicklung festgelegt und in den folgenden Mitosen dieser Zellen beibehalten. Einige der Mechanismen hinter dieser differenzierten Expression werden in der Epigenetik (altgr. *epi ‚jenseits', ‚außerdem'* und *Genetik*) untersucht. Diese Regulationen basieren auf vererbbaren chemischen Strukturanhängseln, die jedoch nicht eine Veränderung der DNA-Sequenz selbst auslösen. Darüber hinaus unterliegen diese Mechanismen einem dynamischen Prozess und sind i. d. R. reversibel (JAENISCH & BIRD, 2003).

Damit Gene im Rahmen der Transkription abgelesen werden können, benötigt die RNA-Polymerase Zugang zu der mit Proteinen (z.B. Histone) verpackten DNA, dem sogenannten Chromatin. Zusätzlich wird zwischen dem Eu- und Heterochromatin unterschieden. Im Heterochromatin liegt die DNA bis auf wenige Ausnahmen stark kondensiert vor und ist folglich für die Polymerase nicht ablesbar. Daher sind die meisten aktiven Gene im bereits entspiralisierten Euchromatin vorzufinden. Jedoch werden auch hier nicht alle Gene auf die gleiche Weise transkribiert und das Euchromatin folglich noch in *aktive* und *nicht-aktive* Bereiche weiter differenziert (JANNING & KNUST, 2004). Wesentliche epigenetische Einflüsse auf den Aktivitätszustand einzelner Gene haben die im Folgenden genauer erläuterten Histonmodifikationen und die DNA-Methylierung.

2.6.1. Histonmodifikationen

Posttranslationale Modifikationen der Histonproteine können die Struktur der DNA so verändern, dass Gene aktiviert oder stummgeschalten (*engl. Gen silencing*) werden. Zu diesen Modifikationen zählen u. a. die Acetylierung, Phosphorylierung und Methylierung der N-terminalen Bereiche eines Histons

(GRAW, 2006). Die Bindung einer Acetylgruppe (-CCH_3) an Lysin-Reste der Histone mittels Histon-Acetyltransferase neutralisiert die positive Ladung der Histone. Daraufhin wird die Bindung zu den negativ geladenen Phosphatresten der DNA gelockert und die RNA-Polymerasen kann die Transkription der Gene starten (JANNING & KNUST, 2004). Im Umkehrschluss wird die Acetylgruppe mit einer Histon-Deacetylase wieder entfernt und die zugrundenliegenden Gene werden inaktiviert (GRAW, 2006). Eine weitere Änderung des Aktivitätszustands wird durch die Phosphorylierung (-PO_3^{2-}) der Histone, v. a. der Aminosäurereste Serin, Threonin und Tyrosin erreicht (MEYERS, 2004). Hierbei wird die Phosphorylgruppe durch Kinasen angehängt und durch Phosphatasen wieder entfernt (OKI et al., 2007). Die Methylierung (-CH_3) findet ebenfalls bevorzugt an den Lysin- sowie an den Arginin-Resten der Histone statt. Durch die Methylierung mittels Histon-Methyltransferasen und Demethylierung durch Histon-Demethylasen wird im Gegensatz zur Acetylierung und Phosphorylierung jedoch nicht die elektrische Ladung verändert. Je nachdem wo sich die Methylgruppe befindet, werden Gene aktiviert oder stummgeschalten. Zusätzlich erhöht sich die Komplexität, da Lysin ein-, zwei- oder dreifach methyliert und Arginin einfach sowie symmetrisch und asymmetrisch di-methyliert werden kann. Einfache Methylierungen sind in der Regel im Euchromatin oder an Enhancern, zwei- und dreifache dagegen im Bereich des Heterochromatins oder der Promotoren anzusiedeln (BANNISTER & KOUZARIDES, 2011). Der Promoter spiegelt die Region der DNA wider, welche die Transkription eines Gens durch die RNA-Polymerase initiiert. Die Aktivität der Promotoren wird wiederrum von spezifischen Enhancern verstärkt (GRAW, 2006). Befindet sich nun zusätzlich zu der Methylierung in der Nähe des Promotors bzw. Enhancers ein acetyliertes Lysin, werden diese aktiviert und in Folge die Genexpression eingeleitet (CALO & WYSOCKA, 2013).

2.6.2. DNA-Methylierung

Die DNA-Methylierung ist eine der meist erforschten epigenetischen Modifikationen bei Menschen, Tieren, Pflanzen und Pilzen (SMITH & MEISSNER, 2013). Hierbei wird der Base Cytosin bzw. dem CpG Dinukleotid

eine Methylgruppe angehängt. Häufig befinden sich diese GC-reichen Regionen ca. 60–100 Basenpaare vor einem Initationscodon eines Gens (GRAW, 2006). Von den ca. 28 Millionen CpGs des humanen Genoms liegen 60–80 % in methylierter Form vor und inaktivieren somit das zugrundeliegende Gen. Weniger als 10 % sind resistent gegenüber dieser Modifikation und befinden sich als CpG-Inseln (ca. 500 – 2.000 bp lang (GRAW, 2006)) nahe den Promotoren von Haushalts- und entwicklungsregulatorischen Genen (SMITH & MEISSNER, 2013). Die exakte Verteilung der methylierten CpGs eines Individuums wird während der Differenzierung der Keimzellen festgelegt. Zuallererst erfolgt eine vollständige Demethylierung im Zuge der embryonalen Zellteilung mit anschließender genomweiter *de novo* Methylierung nach der Implantation der Blastocyste (JAENISCH & BIRD, 2003). Entscheidend für diese Vorgänge sind die Enzyme der Demethylierung, die DNA-Demethylase, und die der *de novo* Methylierung, die DNA-Methyltransferase 1 (DNMT1), DNMT3A und DNMT3B (SMITH & MEISSNER, 2013). Dieser Prozess ist allerdings nicht starr festgelegt und unterliegt einer stetigen Dynamik, u. a. ausgelöst durch verschiedene Umweltfaktoren. Einerseits können Methylgruppen wieder von den methylierten CpG Dinukleotiden entfernt werden und andererseits unmethylierte Dinukleotide methyliert werden (JAENISCH & BIRD, 2003). Basierend auf diesem Hintergrund konnte festgestellt werden, dass mit fortschreitendem Alter eines Individuums vermehrt CpG-Inseln in der Nähe von Tumorsuppressorgenen eine Hypermethylierung aufweisen und eine direkte Korrelation zur Entstehung von Krebszellen besteht. Die Hypermethylierung der CpG-Inseln in Promotoren von Tumorsuppressorgenen führt zu einer Inaktivierung dieser. In Folge werden natürliche Regulationsmechanismen des Zellzyklus und die Apoptose außer Kraft gesetzt, wodurch die Zellteilung nun über das physiologische Maß hinaus stattfinden kann (DAMMANN et al., 2017).

2.7. Kartierungsstudien zu Kalbeverlaufsmerkmalen in Holstein-Friesian Rindern

Zahlreichen Studien gelang es bisher Holstein-spezifische QTL assoziiert mit Fertilitäts- und Kalbemerkmalen auf verschiedensten Chromosomen zu kartieren, darunter BTA1, 3, 5, 6, 7, 8, 12, 18, 20 und 26. Der QTL mit den größten Effekten auf die Merkmale Fertilität und Kalbeverlauf wurde jedoch von mehreren Autoren auf dem Chromosom 18 festgestellt (MA et al., 2019). Zudem war dieser signifikante QTL nach unserem Kenntnisstand bisher ausschließlich in Holstein Rindern bzw. mit HF veredelten Kreuzungslinien nachweisbar. Daher wird im Folgenden nur auf die Studien eingegangen, denen es gelang, den bedeutendsten QTL dieser Rasse auf BTA18 zu kartieren.

Die ersten Studien zur Kartierung von QTL assoziiert mit den Merkmalen Kalbeverlauf und Totgeburt in Holstein-Friesian Rindern erfolgten nach dem Prinzip der Kopplungsanalyse unter Verwendung von Mikrosatellitenmarkern. Für die Kartierung wurde häufig auf lineare Regressionsanalysen zurückgegriffen, erstmals beschrieben von HALEY and KNOTT (1992). Hierbei wurde angenommen, dass sich der QTL zwischen zwei ko-dominaten flankierenden Markern befand. Die relative Lage des QTL in Centimorgan wurde in einer F2-Generation, basierend auf der Kreuzung zweier Inzuchtlinien, anhand der Rekombinationsereignisse zwischen den flankierenden Marker geschätzt. Die Anzahl der verwendeten Mikrosatellitenmarker und dementsprechend die Dichte an Markern auf der bovinen Genkarte variierte hierbei stark zwischen den einzelnen Studien. SCHNABEL et al. (2005) verwendeten zum Beispiel 221 Marker mit einem durchschnittlichen Abstand zwischen den Markern von 15,2 cM, während ASHWELL et al. (2004) 406 Marker mit einem durchschnittlichen Abstand von 7,4 cM in ihren Analysen nutzten. Dementsprechend variierte die Aussagekraft der einzelnen Studien enorm. Trotz der teilweise niedrigen Markerdichte konnten zahlreiche Autoren in verschiedenen Holstein Populationen Quantitative Trait Loci, die mit Kalbemerkmalen assoziiert sind, auf dem Chromosom 18 detektieren. KÜHN et al. (2003) lokalisierten einen QTL assoziiert mit Langlebigkeit, maternalem Kalbeverlauf und Totgeburten in Deutschen Holsteins. In einer Studie von HOLMBERG and ANDERSSON-EKLUND (2006) wurde in der schwedischen Zuchtpopulation von Rot- und

Schwarzbunten Holstein Rindern ein QTL mit maternalen Kalbeeffekten in Verbindung gebracht. THOMASEN et al. (2008) gelang es in dänischen Holstein Populationen pleiotropische Effekte zwischen Schwergeburten und der Größe von Kälbern nachzuweisen. Zudem detektierten SCHNABEL et al. (2005) und KOLBEHDARI et al. (2008) einen QTL für Körperkonstitutionsmerkmale auf dem Chromosom 18 in nordamerikanischen Holsteinlinien. Nichtsdestotrotz konnte in keiner dieser Studien nachgewiesen werden, dass es sich hierbei um ein und denselben QTL handelte. Des Weiteren konnten aufgrund der geringen Markerdichte der Mikrosatellitenmarker die QTL lediglich auf einen relativ großen Bereich lokalisiert werden, wodurch es nicht möglich war, Kandidatengene oder kausale Mutationen zu identifizieren.

Um eine präzisere Kartierung von QTL zu gewährleisten, lösten Single Nukleotid Polymorphismen die Mikrosatelliten als Marker in den folgenden Jahren vollständig ab (siehe Kapitel 2.3.2.1). Inzwischen werden zur Identifikation von QTL vornehmlich die genomweite Assoziationsstudien (GWAS) und die kombinierte Kopplungsungleichgewichts- und Kopplungsanalyse (cLDLA) angewendet. Im Rahmen zahlreicher Studien konnte mit Hilfe von SNP Markern ein signifikanter QTL auf einen Bereich zwischen 50 und 60 Mbp auf dem Chromosom 18 eingegrenzt werden (COLE et al., 2009; COLE et al., 2011; SAHANA et al., 2011; MAO et al., 2016; PARKER GADDIS et al., 2016; MÜLLER et al., 2017). Darüber hinaus konnte in der Studie von MÜLLER et al. (2017) erstmals im Bereich des QTL ein kausaler Haplotyp mit Hilfe von HD-Markern in einem Bereich zwischen 57.941.736 bp und 58.442.683 bp (ARS-UCD1.2) lokalisiert werden. Der kausale Haplotyp assoziiert mit den Merkmalen paternaler Kalbeverlauf und Totgeburt trat mit einer Frequenz von 13,3 % in Deutschen Holsteins auf. COLE et al. (2009) gelang es außerdem einen SNP (rs109478645) an der Position 57.137.302 bp (auf dem ARS-UCD1.2) assoziiert mit den Merkmalen maternaler bzw. paternaler Kalbeverlauf, Beckenbreite, Größe, Stärke und Körpertiefe zu identifizieren. Lediglich 48 kb von dem SNP von COLE et al. (2009) entfernt, detektierten MAO et al. (2016) an der Position 57.089.460 bp (ARS-UCD1.2) den SNP rs136283363 mit einer signifikanten Assoziation zu paternalen Kalbemerkmalen. Aufgrund des geringen Abstandes wurde in Folge ein hohes Kopplungsungleichgewischt (0,89) zwischen diesen Markern

angenommen. Darüber hinaus lokalisierten COLE et al. (2009) den signifikanten SNP rs109478645 an der Position eines Introns des *sialic acid binding IG-like lectin* (*SIGLEC*) *5* Gens (nach der neuen Nomenklatur *SIGLEC-13* (57.136.157 – 57.142.779 bp)). Weitere Studien, darunter SEIDENSPINNER et al. (2009), PURFIELD et al. (2015) und MAO et al. (2016), konnten zudem das *SIGLEC13*-Gen als vielversprechendes Kandidatengen bestätigen. Über das derselben Genfamilie zugehörige humane *SIGLEC-6* Gen konnte ein Zusammenhang zur Geburtseinleitung hergestellt werden. Die Genprodukte des *SIGLEC-6* Gens werden in der humanen Plazenta exprimiert und sind dazu in der Lage Leptin zu binden. Dieses Leptinbindungsvermögen stellt einen Mechanismus der Geburtseinleitung dar und trägt damit auch zur Dauer der Trächtigkeit bei (BRINKMAN-VAN DER LINDEN et al., 2007). Sind nun die Genprodukte des *SIGLEC-6* bzw. *13* oder deren Expression aufgrund einer Mutation verändert, kann dies in Folge zu einem verzögerten Geburtsbeginn durch die mangelnde Bindung an Leptin führen. Des Weiteren konnte eine unerwünschte Korrelation zwischen Kalbe- und Körperkonstitutionsmerkmalen nachgewiesen werden. Mit einer verstärkten Selektion auf Körpergröße und Gewicht führte dies folglich zu größeren und schweren Kälbern zum Zeitpunkt der Geburt. Im Gegensatz dazu korrelierte die Zunahme der Körperkonstitutionsmerkmale nicht mit einer Erweiterung des Beckens des Muttertiers. Die Inkompatibilität zwischen größeren Kälbern bei vergleichsweise kleinem Beckendurchmesser des Muttertiers erhöhte somit zusätzlich die Anzahl an Schwer- und Totgeburten (HANSEN et al., 2004; COLE et al., 2009). Aufgrund der hohen Komplexität sowie niedrigen Heritabilität von Kalbemerkmale konnten bisher jedoch in keiner Studie das *SIGLEC-13* Gen als eindeutiges Kandidatengene bestätigt bzw. dessen kausale Mutation(en) identifiziert werden (MA et al., 2019). FANG et al. (2019) und PURFIELD et al. (2020) detektierten in nordamerikanischen bzw. irischen Holstein Zuchtlinien im Bereich des signifikanten QTL einen weiteren SNP (rs31577268) assoziiert mit Kalbeverlaufsmerkmalen an der Position 57.816.137 bp (ARS-UCD1.2). Der SNP rs31577268 konnte zudem in der Nähe des *Zink-Finger-Proteins 613*-Gens (*ZNF613*) lokalisiert werden, wodurch ein weiteres aussichtsvolles Kandidatengen identifiziert wurde. Jedoch ist auch im Falle des *ZNF613*-Gens weitere Forschung notwendig, um einen kausalen Mechanismus mit Effekten auf dem Kalbekomplex bestätigen zu können. Neben der genomweiten Assoziationsstudie untersuchten FANG

et al. (2019) die DNA-Methylierung in Spermienproben von Holstein Bullen. Im Zuge dessen konnte im zweiten Intron des *ZNF613*-Gens eine differentiell methylierte Region (DMR) identifiziert werden, die mit den Merkmalen Trächtigkeitsdauer, paternaler Kalbeverlauf, Körpertiefe und Konzeptionsrate assoziiert werden konnte. Die Autoren stellten fest, dass Tiere mit einer höheren Methylierungsrate im zweiten Intron zwar eine höhere Konzeptionsrate aufwiesen, zusätzlich aber auch eine verlängerte Trächtigkeitsdauer, größere Kälber und mehr Schwergeburten beobachtet werden konnten. FANG et al. (2019) beschrieben allerdings nur vage die verwendete Methode zur Feststellung von differentiell methylierten Regionen und gaben zudem einen relativ großen Bereich an (ca. 30 kb) in der sich die mögliche kausale Methylierung befand. Aus diesem Grund bedarf es weiterer Studien, um einen tatsächlichen Zusammenhang zwischen der Methylierung möglicher kausaler Gene wie dem *ZNF613*-Gen und dem Kalbekomplex in HF Rindern zu bestätigen.

Ziel dieser Arbeit ist es daher, den Bereich des signifikanten QTL erneut zu analysieren, um potenzielle Kandidatengene und deren kausale Mutation(en) zu identifizieren. Nur durch die Entschlüsselung des signifikanten QTL ist es möglich, eine anhaltend gute Fruchtbarkeit sowie leichte Kalbungen in der bedeutendsten Milchviehrasse der Welt zu gewährleisten.

3. Material und Methoden

3.1. Material

3.1.1. Zusammenstellung des Tiersets und Auswahl der Phänotypen

Das von MÜLLER et al. (2017) generierte Tierset beinhaltete 2.572 töchtergeprüfte Holstein Bullen. Für diese Studie wurde dieses Probenset auf 2.697 Tiere erweitert, darunter 2.525 Bullen und 172 Kühe. Zu beachten ist, dass von 47 Bullen aus dem Datensatz von MÜLLER et al. (2017) keine aktuellen Zuchtwerte (ZWS April 2019) vorlagen und diese in Folge nicht in das aktuelle Tierset übernommen wurden. Die vorliegenden Proben stammten aus Projekten von Prof. Dr. Georg Thaller, Leiter der Arbeitsgruppe Tierzucht und Haustiergenetik am Institut für Tierzucht und Tierhaltung der Agrar- und Ernährungswissenschaftlichen Fakultät der Christian-Albrechts-Universität zu Kiel, Ph.D. Lilian Gehrke, wissenschaftliche Mitarbeiterin der Vereinigten Informationssysteme Tierhaltung w.V. (VIT) in Verden und aus den vorhanden internen Datensätzen der eigenen Arbeitsgruppe (AG Populationsgenomik) (THALLER, 2011; MEDUGORAC et al., 2012; GEHRKE et al., 2020).

Alle Analysen erfolgten ausschließlich auf dem Chromosom 18 mit dem Hauptaugenmerk auf das Merkmal paternaler Kalbeverlauf, da dieses in der Studie von MÜLLER et al. (2017) die signifikanteste Assoziation zu dem zugrundeliegenden QTL aufwies. Der untersuchte Phänotyp spiegelte den relativen Zuchtwert des Merkmals paternaler Kalbeverlauf eines jeden Tieres des Datensets wider. Die verwendeten Zuchtwerte stammten aus der offiziellen Zuchtwertschätzung des VIT in Verden veröffentlicht im April 2019. In der folgenden Tabelle 3 ist eine deskriptive Statistik der verwendeten Zuchtwerte aufgeführt. Es wurde angenommen, dass bei der Zucht mit Tieren, die einen niedrigen relativen Zuchtwerten für das Merkmal pKV aufwiesen, die Wahrscheinlichkeit einer Schwergeburt höher ist, als dies bei Tieren mit hohen Zuchtwerten der Fall wäre.

Zur Erweiterung des Datensatzes und Validierung der Ergebnisse wurden aus dem Sequence Read Archive (SRA) des *National Center for Biotechnology* (NCBI) freiverfügbare Ganzgenomsequenzdaten von 21 Holstein Tieren, davon 18 männliche und drei weibliche, heruntergeladen. Die ausgewählten Genome wurden mittels paired Illumina Short-Read Sequenzierungstechnik sequenziert und stammten aus verschiedenen Projekten, die sich nicht näher

mit Fruchtbarkeitsmerkmalen in HF Rindern beschäftigten. Eine Auflistung der verwendeten Proben und den zugehörigen Zugangsnummern der Projekte befindet sich im Anhang 1.

Darüber hinaus wurde das zusammengestellte Tierset für die weiteren Analysen mit den Sequenzdaten weiterer Nicht-Holstein Rassen komplementiert. In die Mappingstudie wurden zusätzlich die ONT sequenzierten Genome eines Kärntner Blondvieh aus dem internen Datensatz und eines Fleckvieh Bullen, publiziert in der Studie von GEHRKE et al. (2020), sowie aus dem NCBI SRA paired Illumina Short-Read Ganzgenomsequenzen der Rassen Hereford (SRR8324584) und Brahman (SRR2016745) aufgenommen. Die Sequenzen aus dem SRA sind unter den folgenden Projektnummern in der Bio-Projekt-Datenbank des NCBI einsehbar: PRJNA494431 (Hereford) und PRJNA277147 (Brahman).

Tabelle 3: Deskriptive Statistik der verwendeten Zuchtwerte des Merkmals paternaler Kalbeverlauf

Mittelwert	Minimum	Maximum	Modalwert	Median	Varianz	Standardabweichung
97,84	49,14	130,51	96,23	98,01	86,82	9,32

3.2. Methoden

3.2.1. Verwendete Programme

Die für diese Arbeit notwendige Software sind in der nachfolgenden Tabelle 4 nach ihrem Verwendungszweck aufgelistet.

Tabelle 4: Verwendete Software

Programmname	Verwendungszweck und Quelle
ASReml	Varianzkomponentenanalyse GILMOUR et al. (2009) https://www.vsni.co.uk/software/asreml/
Beagle 5.0	Haplotypisierung und Imputation von Genotypen BROWNING and BROWNING (2009); BROWNING et al. (2018) https://faculty.washington.edu/browning/beagle/b5_0.html
BreakDancer v1.4.5	Identifikation struktureller Varianten in Illumina Short-Read Sequenzen CHEN et al. (2009) https://github.com/genome/breakdancer
BWA 0.7.17	Mapping Illumina Short-Reads gegen das Referenzgenom LI and DURBIN (2009) http://bio-bwa.sourceforge.net/
Circos v0.52	Visuelle Darstellung segmentaler Duplikationen KRZYWINSKI et al. (2009) http://circos.ca/
FASTQC v0.11.8	Konvertierung in das FASTQ Format und Qualitätskontrolle der Illumina Short-Read Sequenzen BABRAHAM INSTITUTE (2017) http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
GATK	Genotypenbestimmung und Identifikation von SNP und Indels MCKENNA et al. (2010) https://gatk.broadinstitute.org/hc/en-us
Guppy 3.2.2	Base-Calling PAYNE et al. (2020) https://nanoporetech.com/nanopore-sequencing-data-analysis
IGV 2.6.3	Visualisierung der gemappten Sequenzen THORVALDSDOTTIR et al. (2013) https://software.broadinstitute.org/software/igv/download
Lumpy v0.3.1	Identifikation struktureller Varianten in Illumina Short-Read Sequenzen LAYER et al. (2014) https://github.com/arq5x/lumpy-sv
Manta v1.6.0	Identifikation struktureller Varianten in Illumina Short-Read Sequenzen CHEN et al. (2016) https://github.com/Illumina/manta
MegaBLAST	Selbstkartierung des Chromosom 18 ZHANG et al. (2000) https://github.com/gpertea/mgblast
Microsoft Visual Studio 2012	Erstellung von Programmen zur Datenvorbereitung, -analyse und -auswertung https://visualstudio.microsoft.com/de/
Minimap2-2.17	Mapping der ONT Long-Reads gegen das Referenzgenom LI (2018) https://github.com/lh3/minimap2/releases

NanoVar v1.3.8	Identifikation struktureller Varianten in ONT Long-Read Sequenzen THAM et al. (2020) https://github.com/benoukraflab/NanoVar
NCBI Genome Remapping Service	Konvertierung der Positionen unterschiedlicher Genome NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (2018) https://www.ncbi.nlm.nih.gov/genome/tools/remap
Picard 2.20.7	Entfernung möglicher Duplikate in den Sequenzproben BROAD INSTITUTE (2019) https://broadinstitute.github.io/picard/
Porechop v0.2.4	Adapter Trimming der Sequenzen WICK (2018) https://github.com/rrwick/Porechop
R 3.6.1	Datenverarbeitung R CORE TEAM (2019) https://www.r-project.org/
Racon 1.3.3	Qualitätskontrolle im Rahmen *de novo* Assembly Techniken VASER et al. (2017) https://github.com/isovic/racon
SAMtools 1.9	Datenbearbeitung von SAM-/BAM-/VCF-Dateien LI et al. (2009) https://sourceforge.net/projects/samtools/files/samtools/1.9/
SDDetector v0.2	Identifikation segmentaler Duplikationen DALLERY et al. (2017) https://github.com/nlapalu/SDDetector
Sickle 1.33	Trimming FASTQ-Files JOSHI and FASS (2011) https://github.com/najoshi/sickle
SRA-Toolkit 2.9.6	Download frei verfügbarer Sequenzen NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI) (2018b) https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc/
Wtdbg2 2.5	Assembler RUAN and LI (2019) https://github.com/ruanjue/wtdbg2

3.2.2. Verwendete Datenbanken

Ein Überblick über die angewendeten Datenbanken dieser Arbeit ist in Tabelle 5 zu finden.

Tabelle 5: Verwendete Datenbanken

Datenbank	Quelle
BGVD	Überprüfung potenzieller Kandidaten-SNPs CHEN et al. (2020) http://animal.nwsuaf.edu.cn/code/index.php/BosVar
Ensembl	Ensembl BLAST und DGVa zur Kontrolle struktureller Varianten CUNNINGHAM et al. (2018) http://www.ensembl.org/index.html
Gendatenbank des NCBI	Überprüfung potentieller Kandidatengene https://www.ncbi.nlm.nih.gov/gene/
Microsoft SQL-Server 2014	Interne Tierdatenbank (u.a. Geno- und Haplotypen, Einzeltier- und Abstammungsinformationen) https://www.microsoft.com/de-de/sql-server/
PubMed®	Literaturrecherche https://www.ncbi.nlm.nih.gov/pubmed/
Sequence Read Archiv (NCBI)	Download der freiverfügbare Ganzgenomsequenzen NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI) (2018a) https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement

3.2.3. Kombinierte Kopplungsungleichgewichts- und Kopplungsanalyse

3.2.3.1. Genotypisierung

Alle 2.697 Proben des Tiersets wurden bereits in vorherigen Projekten genotypisiert. Dazu wurde sowohl der BovineSNP50 BeadChip als auch bei 256 Tieren der BovineHD BeadChip verwendet. Der 50K BeadChip in Version 1 enthält 54.001 SNPs, als Version 2 54.609 SNPs und der HD-Chip beinhaltet 777.962 SNPs (Illumina, San Diego, USA). Die Markerpositionen wurden mit Hilfe des NCBI GENOME REMAPPING SERVICE (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2018) von dem in MÜLLER et al. (2017) verwendete *Bos taurus* UMD 3.1.1 Referenzgenom (ZIMIN et al., 2009) auf das aktuelle ARS-UCD1.2 Assembly (ROSEN et al., 2020) konvertiert.

Um möglichst exakte Ergebnisse zu erhalten, durchliefen alle Marker eine spezifische Qualitätskontrolle. All jene SNPs, die ein oder mehr der folgenden Kriterien erfüllten, wurden von den folgenden Analysen ausgeschlossen:

1. SNPs, die in über 5 % der Tiere nicht erfolgreich genotypisiert werden konnten.
2. SNPs, deren Position auf der ARS-UCD1.2 Genkarte nicht eindeutig bzw. überhaupt nicht zuordenbar war.
3. SNPs mit einer „minor allele frequency" (MAF) von unter 2,5 %.
4. SNPs, die zu vermehrten Konflikten aufgrund der Abstammung verwandter Tiere führten (d.h. Mendelian Fehlerrate von mehr als 0,2 %).
5. Alle Marker, die nicht auf dem Chromosom 18 lokalisiert waren.

Nach Durchführung des Filterprozesses konnten insgesamt 1.144 Marker in das Studiendesign aufgenommen werden.

3.2.3.2. Erweiterung der Genotypendatenbank

Zur Erweiterung des Datensatzes wurden zusätzlich die SNP-Genotypen aus frei verfügbaren Ganzgenomsequenzen (engl. whole-genome sequencing – WGS) von 21 Holstein Tieren bestimmt. Die paired Illumina sequenzierten Daten standen über das NCBI Sequence Read Archiv zum Download bereit

(siehe Anhang 1). Für die Genotypenbestimmung wurden verschiedene Programme des *GENOME ANALYSIS TOOLKIT* (MCKENNA et al., 2010) verwendet, darunter die Programme *HAPLOTYPECALLER* (*HC*), *GENOTYPEGVCF* und *VARIANTFILTRATION*.

Zur Verwendung des *HAPLOTYPECALLER* bedarf es Sequenzdaten im BAM (*Binary Alignment Map*) Format. Die heruntergeladenen Rohdaten der 21 Proben wurden dafür zunächst mit Hilfe des Programms FASTQC v0.11.8 (BABRAHAM INSTITUTE, 2017) in FASTQ-Files umgewandelt und anschließend mit Hilfe des *BURROWS-WHEELER ALIGNER* Version 0.7.17 (LI & DURBIN, 2009) gegen das ARS-UCD1.2 Referenzgenom gemappt. Die einzelnen Schritte des Mappings und die Erstellung der BAM Files sind im Kapitel 3.2.5.2 näher erläutert. Mit Hilfe der generierten BAM Dateien konnte nun die Genotypenbestimmung gestartet werden. POPLIN et al. (2018) unterteilten die Vorgehensweise des HC in vier Schritte:

(1) Identifikation sogenannter *aktiver Regionen*, d.h. der Regionen im Genom, die ein Alternativallel im Vergleich zum Referenzallel aufweisen könnten. Zu beachten ist, dass in dieser Studie im ersten Schritt die Option „includeNonVariantSites" hinzugefügt wurde, um für jede Position im Genom den Genotyp unabhängig davon zu bestimmen, ob das Allel an dieser Position vom Referenzallel abwich oder nicht.

(2) Ermittlung aller potentieller Haplotypen einer Probe mit dem Ziel das reale physikalische DNA-Fragment zu rekonstruieren, das die *aktive Region* bzw. in dieser Studie die jeweilige Position (d.h. Positionen für SNPs auf dem 50K und HD Chip) beinhaltet. Hierfür werden mit Hilfe von Referenz-basiertem Assembling alle Reads mit der *aktiven Region* bzw. Position gegen die Sequenz eines Referenzgenoms gemappt. In Folge werden nicht nur die Haplotypen für das Allel der *aktiven Region* bzw. der Position in jedem Read, sondern auch die Haplotypen für die gesamte Sequenzlänge überlappender Reads bestimmt.

(3) Schätzung der Haplotypen-Wahrscheinlichkeiten pro Position mittels einem *paired Hidden Markov Model* Algorithmus. Hierbei wird jeder Read gegen jeden möglichen Haplotypen aus Schritt (2) einschließlich dem Haplotypen des Referenzallels gemappt. Daraus ergibt sich für jeden potentiellen Haplotypen eine Wahrscheinlichkeit, wie häufig dieser in dem jeweiligen Read auftritt.

(4) Bestimmung des Genotyps der *aktiven Region* bzw. Position. Anhand der im Schritt (3) ermittelten Wahrscheinlichkeiten für jeden potentiellen Haplotypen pro Read wird nun der wahrscheinlichste Genotyp mit Hilfe des Satzes von Bayes für Wahrscheinlichkeitsberechnung geschätzt.

Das Programm *VARIANTFILTRATION* wurde zur Qualitätskontrolle der abgeleiteten Genotypen verwendet. Der Standardfilter des Programms umfasste sämtliche Kriterien, die falsche Ergebnisse aufgrund eines Bias von Referenz- und Alternativallel reduzieren (z.B. unterschiedliche Genauigkeit des Mappings von Referenz- und Alternativallel) (CHANDRAN, 2016). Zusätzlich wurden die *Parameter Depth of Coverage* (DP) und Qualität der geschätzten Genotypen (GQ) mit in die Qualitätskontrolle aufgenommen.

Die Genotypenqualität (GQ) wird mit Hilfe der normalisierten Phred-Scores der drei möglichen Genotypen homozygot Referenzallel (0/0), heterozygot Referenz- und Alternativallel (0/1) und homozygot Alternativallel (1/1) (VAN DER AUWERA, 2017) ermittelt. Der Phred-Score oder Q-Score ist eine der am häufigsten verwendeten Maßeinheiten zur Bestimmung der Qualität einer Sequenzierung. Mit diesem Wert wird die Wahrscheinlichkeit geschätzt, dass die sequenzierte Base falsch ist (ILLUMINA INC., 2011). Der Qualitätsparamater (Q) wird mit der folgenden Formel definiert:

$$Q = -10\, log_{10}\, P$$

wobei P die geschätzte Fehlerwahrscheinlichkeit der abgerufenen Base ist. Das bedeutet, bei einer Fehlerwahrscheinlichkeit der Base von $P = 0{,}001$ ergibt dies einen Q-Score von 30 (EWING & GREEN, 1998). Das *VARIANTFILTRATION* Programm normalisiert nun die Q-Scores der drei Genotypen, sodass der Genotyp mit der höchsten Fehlerwahrscheinlichkeit (d.h. der Genotyp mit dem niedrigsten Q-Score) 0 ist. Die Genotypenqualität einer Position entspricht nun immer dem zweitniedrigsten Q-Score (z.B. GQ = 20, bei den drei normalisierten Q-Scores 40, 20 und 0). Des Weiteren spiegelt die Genotypenqualität die Differenz zwischen den Likelihoods der zwei wahrscheinlichsten Genotypen wider. Das bedeutet, umso höher die Differenz zwischen den beiden Genotypen ist, desto weniger wahrscheinlich ist der zweitwahrscheinlichste Genotyp korrekt. In dieser Studie wurden nur Genotypen mit einer GQ von mehr als 22 berücksichtigt (VAN DER AUWERA, 2017).

Die Depth of Coverage spiegelte die Gesamtanzahl der Reads wieder, denen ein Allel zugeordnet werden konnte und das die Standardkriterien erfüllte (VAN DER AUWERA, 2017). Hierfür wurde ein Mindestmaß von sieben Reads festgelegt, da bei einer Binomialverteilung die Wahrscheinlichkeit, dass das vorhandene Allel nicht erkannt wurde, bei unter 0,01 ($0{,}5^7 = 0{,}0078$) lag.

3.2.3.3. Haplotypisierung und Imputation

Zur Durchführung der kombinierten Kopplungsungleichgewichts- und Kopplungsanalyse (cLDLA) bedarf es der differenzierten (phased) Haplotypen jedes Individuum. Ei- und Samenzelle enthalten jeweils einen vollständigen haploiden Chromosomensatz. Folglich ist der Haplotyp die Gesamtheit aller Allele eines Chromosoms und wird von jeweils einem Elternteil vererbt. Im Zuge der Befruchtung verschmelzen anschließend die beiden Haplotypen in der Zygote zu einem diploiden Chromosomensatz (2n) (JANNING & KNUST, 2004). Um nun die beiden Haplotypenpaare aus dem diploiden Chromosomensatz zu differenzieren (Phasing), wurden diese anhand der einzelnen Genotypen und einem diploiden Hidden Markov Model (HMM) mit Hilfe des Programms *BEAGLE* 5.0 geschätzt (BROWNING & BROWNING, 2007; BROWNING et al., 2018). Im ersten Schritt erfolgte eine Gruppierung aller Haplotypen für jeden Marker im Genom, abhängig von den Allelen an dieser Position (BROWNING & BROWNING, 2009). Anhand des lokalisierten Haplotypengruppen-Modells konnte das diploide HMM abgeleitet werden und in Abhängigkeit der individuellen Genotypen die beiden differenzierten Haplotypen ermittelt werden. Im finalen Schritt des Phasing-Algorithmus wurde der wahrscheinlichste mütterliche als auch väterliche Haplotyp ausgewählt (BROWNING & BROWNING, 2007).

Um eine akkurate Haplotypisierung durchführen zu können, bedarf es an möglichst jeder Markerposition einen Genotyp. Die Ableitung fehlender Genotypen, die sogenannte Imputation, wurde ebenfalls mit dem Programm *BEAGLE* 5.0 durchgeführt. Das Grundprinzip basiert auf der sogenannten *identity by descent* (IBD) Wahrscheinlichkeit. Diese liegt vor, wenn zwei Chromosomenabschnitte von ein und demselben Vorfahren abstammen und kaum Rekombinationsereignisse auftraten. IBD Segmente weisen daher bis auf einzelne mutierte Positionen die gleichen Allelsequenzen auf. Die

gegebenen Genotypen des Datensatzes konnten genutzt werden, um IBD-Abschnitte zu identifizieren, indem die Haplotypen des Individuums mit denen der Referenz verglichen wurden. Anschließend wurden nicht genotypisierte IBD-Bereiche mit den IBD-Segmenten des Referenzhaplotypen ergänzt. Um eine ausreichend hohe Sicherheit zu gewährleisten, wurde im Vorfeld mit Hilfe eines HMM die Wahrscheinlichkeit für jedes mögliche Allel der nicht genotypisierten Marker ermittelt (BROWNING et al., 2018).

Die Ausführung der Haplotypisierung und Imputation mit *BEAGLE* 5.0 erfolgte in drei separaten Durchläufen und unterschiedlichen Datensets. Im ersten Datensatz wurden all jene Tiere berücksichtigt, die zuvor mit dem 50K BeadChip genotypisiert wurden. Dieser Datensatz umfasste zusätzlich Tiere mit HD Genotypen, reduziert auf die Marker des 50K BeadChips. Der zweite Datensatz beinhaltete nur die Tiere, welche mit dem BovineHD BeadChip genotypisiert wurden. In beiden Durchgängen wurden alle verfügbaren Genotypen verwendet, inklusive der Individuen, die ansonsten nicht weiter in der cLDLA verwendet wurden. Die letzte Haplotypisierung und Imputation erfolgte mit den Tieren aus dem zweiten Datensatz ergänzt um die Genotypen der 21 Illumina sequenzierten Holstein Proben (siehe Kapitel 3.2.3.2).

3.2.3.4. Korrektur der Verwandtschaftsbeziehungen zwischen Individuen

Bei der Kartierung von QTL in Nutztierpopulationen treten aufgrund der Halb- (z.B. Rind) bzw. Vollgeschwister-Strukturen (z.B. Schwein) verwandtschaftsabhängige falsch positive Ergebnisse auf, wie bereits in Kapitel 2.3.3.2 näher beschrieben. Aus diesem Grund sollten bei der Schätzung von polygenen Effekten möglichst präzise Verwandtschaftskoeffizienten zwischen allen Tieren der Kartierungspopulation als Variablen in der cLDLA berücksichtigt und somit das Auftreten falscher Ergebnisse reduziert werden (GODDARD & HAYES, 2009). Die vereinheitlichte additiv-genetische Verwandtschaft (*Unified Additive Relationships*, UAR) wurden für alle Tiere der Kartierungspopulation geschätzt und im Anschluss eine Matrix daraus gebildet. Entscheidend für die Schätzung ist die Wahrscheinlichkeit von Herkunftsidentitäten (*identity by descent*, IBD) zwischen Allelen in zwei verschiedenen Gameten unter

Berücksichtigung einer Basis- bzw. Referenzpopulation. Bei herkunftsgleichen Allelen bzw. Haplotypen wurde davon ausgegangen, dass diese von einem gemeinsamen Vorfahren abstammten. Je nachdem, ob sich die Allele im gleichen diploiden Tier befanden, konnte eine Aussage über den Inzuchtkoeffizienten getroffen werden bzw. bei verschiedenen Individuen über deren Verwandtschaftsbeziehung zueinander. Zwei Allele oder DNA-Segmente sind zustandsgleich (*identity by state*, IBS), wenn sie identische Allele bzw. Nukleotidsequenzen in diesem Segment aufweisen. Das bedeutet, DNA-Abschnitte, die von einem gemeinsamen Vorfahren abstammen (IBD-Segmente), sind per Definition auch immer IBS. Andersherum sind jedoch nicht alle IBS Segmente zwangsläufig auch IBD. Verschiedene Individuen können identische DNA-Regionen aufweisen, die durch dieselben Mutationen oder Rekombinationen und nicht aufgrund eines gemeinsamen Vorfahren entstanden sind (POWELL et al., 2010). Die genomweite Verwandtschaftsmatrix zwischen allen Tieren der Kartierungspopulation wurde mit dem R-Paket *snpReady* geschätzt. Im Anschluss erfolgte die Berechnung einer generalisierten Inversen mit dem Programm *ginverse* (Karin Meyer, University New England, Australien) aus der Verwandtschaftsmatrix. Die generalisierte Inverse wurde in die nachfolgende Varianzkomponentenanalyse zur Schätzung zufälliger polygener Effekte miteinbezogen. Um falsch positive Ergebnisse aufgrund von Familienstrukturen und Populationsstratifikationen zu vermeiden, wurden die zufällig polygenen Effekte zuvor korrigiert.

3.2.3.5. Locus IBD und Diplotypen-Verwandtschaftsmatrix

Wie in Kapitel 3.2.3.4 erwähnt, wird zur Schätzung genomweiter polygener Effekte eine genomweite Verwandtschaftsmatrix berücksichtigt. Analog ist zur Schätzung von QTL, d.h. den lokalen genetischen Effekten auf ein quantitatives Merkmal, eine lokale Verwandtschaftsmatrix notwendig. Zur Kartierung der QTL auf einem Chromosom bzw. im Genom sollten die lokalen Verwandtschaftsmatrizen für alle Loci zwischen den beobachteten Markern abgleitet werden. Zu diesem Zweck wurde eine Locus IBD Matrix für jedes Markerintervall nach der folgenden Vorgehensweise geschätzt: MEUWISSEN and GODDARD (2001) zeigten, dass zur Schätzung der IBD-

Wahrscheinlichkeiten die Marker-Haplotypen eines beliebigen Locus zwischen zwei Markern verwendet werden können. Dabei erwies sich der Mittelpunkt zwischen zwei benachbarten Markern als besonders repräsentativ für das jeweilige Markerintervall. Zur Schätzung der Locus IBD sind mindestens die Haplotypen von zwei Markerloci notwendig, wobei mit der Anzahl (*N*) an verwendeten Loci zwar einerseits die Genauigkeit der Kartierung steigt, allerdings auch der damit einhergehende rechnerische Aufwand. In dieser Studie wurden die Haplotypen von 40 benachbarten SNPs betrachtet und die IBD am informativsten Punkt des Intervalls, d.h. an der Position zwischen den Markern 20 und 21, abgeleitet. Dieses 40 SNP-Markerfenster wurde nun entlang des gesamten Chromosoms bewegt, um eine Locus-IBD Matrix an jedem Mittelpunkt aller Markerintervalle zu schätzen. Da Individuen sowohl einen paternalen als auch einen maternalen Haplotypen aufweisen, ergaben sich somit an jedem Fenstermittelpunkt vier verschiedene IBD-Wahrscheinlichkeiten für jedes Tier-Paar (*i-j*): paternal(*i*)-paternal(*j*), paternal(*i*)-maternal(*j*), maternal(*i*)-paternal(*j*), und maternal(*i*)-maternal(*j*). Um aus diesen vier Haplotypen IBD eine Diplotypen-Verwandtschaftsmatrix (*diplotype relationship matrix*, $\mathbf{D_{RM}}$) zwischen Tier *i* und Tier *j* zu berechnen, wurde, wie von LEE and VAN DER WERF (2006) beschrieben, die Summe der vier Haplotypen IBD durch zwei dividiert. Mit den erhaltenen Diplotypen-Verwandtschaftsmatrizen konnten in der folgenden Varianzkomponentenanalyse die zufälligen QTL-Effekte an den jeweiligen Fenstermittelpunkten ermittelt werden.

3.2.3.6. Varianzkomponentenanalyse

Ziel der kombinierten Kopplungsungleichgewichts- und Kopplungsanalyse ist es einen möglichst kurzen Markerabschnitt zu identifizieren, der mit der größten Wahrscheinlichkeit einen Locus mit signifikanten Effekten auf das zu untersuchende Merkmal enthält (MEUWISSEN & GODDARD, 2000). Zu diesem Zweck wurden in einer Varianzkomponentenanalyse nach MEUWISSEN et al. (2002) sowohl die Informationen des Kopplungsungleichgewichts in Form der Diplotypen-Verwandtschaftsmatrix ($\mathbf{D_{RM}}$) als auch die Kopplungsinformationen der Haplotypenrekonstruktion unter Berücksichtigung der Pedigreedaten miteinbezogen.

Die Varianzkomponentenanalyse erfolgte mit dem Programm *ASREML* (GILMOUR et al., 2009) und dem folgenden gemischten linearen Modell:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_1 u} + \mathbf{Z_2 q} + \mathbf{e}$$

mit den folgenden Variablen:

y = Vektor des Phänotyps, d.h. relativer Zuchtwert des Merkmals paternaler Kalbeverlauf

β = Vektor der fixen Effekte

u = Vektor der zufällig polygenen Effekte

q = Vektor der zufällig additiv-genetischen QTL-Effekte

e = Vektor der zufälligen Resteffekte

X, Z_1, Z_2 = Designmatrizen

Der Vektor **y** gab hierbei das phänotypische Merkmal Kalbeverlauf in Form des relativen Zuchtwertes paternaler Kalbeverlauf (pKV) an. Im Vektor **β** wurden die fixen Effekte wie Geschlecht und der allgemeine Mittelwert μ berücksichtigt. Von den drei Vektoren **u, q** und **e** wurde angenommen, dass diese nicht korreliert sind und normalverteilt mit einem Mittelwert 0. Der Vektor **u** repräsentierte die zufällig auftretenden polygenen Effekte eines jeden Tieres, wobei $\mathbf{u} \sim N(0, \mathbf{G}\sigma_u^2)$ ist und **G** die genomweite Verwandtschaft (UAR-Matrix) zwischen allen Tiere der Kartierungspopulation angibt. Der Vektor **q** steht für die zufällig additiven-genetischen Effekte eines QTL mit $\mathbf{q} \sim N(0, \mathbf{D_{RM}}_i\sigma_q^2)$. $\mathbf{D_{RM}}_i$ spiegelte die $\mathbf{D_{RM}}$ Matrix an jedem Fenstermittelpunkt *i* des Chromosoms wieder. Die restlichen zufälligen Effekte wurden mit dem Vektor $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ aufgegriffen, wobei **I** die Identitätsmatrix ist. Die Berechnung der Varianzen der einzelnen Vektoren (σ_u^2, σ_q^2, σ_e^2) wurde mit *ASREML* durchgeführt. Eine Verknüpfung der beobachteten Werte mit den fixen und zufälligen Effekten erfolgte durch die Designmatrizen **X**, $\mathbf{Z_1}$, $\mathbf{Z_2}$.

3.2.3.7. Likelihood-Ratio Teststatistik

Im letzten Schritt der QTL Kartierung wurde der Likelihood-Ratio Test (LRT) an jedem Fenstermittelpunkt des getesteten Chromosoms durchgeführt. Mit Hilfe der LRT können zwei Modelle miteinander verglichen werden, die sich in nur einem Parameter unterscheiden. Mit der Likelihood-Ratio wird getestet, ob das Modell mit einer zusätzlichen Variablen (d.h. mit Vektor **q**) die Daten besser erklärt als das Standardmodell (d.h. ohne die zufälligen QTL-Effekte).

In dem Modell der Nullhypothese ($logL(H_0)$) waren keine QTL-Effekte enthalten, während in der Alternativhypothese ($logL(H_1)$) die QTL-Effekte berücksichtigt wurden. Die beide Likelihoods wurden ebenfalls mit dem Programm *ASREML* berechnet und die LRT-Werte mit der nachstehenden Formel berechnet:

$$LRT = -2 \times \left(logL(H_0) - logL(H_1)\right)$$

Die LRT-Statistik folgt einer X^2-Verteilung mit einem Freiheitsgrad (HEUVEN et al., 2005). Jener LRT mit dem höchsten Wert gab die wahrscheinlichste Position des QTL an. Zur Visualisierung der errechneten LRT-Kurve wurden die LRT-Werte von jedem Fenstermittelpunkt am Ende in das Programm *MICROSOFT POWER POINT* überführt.

3.2.3.8. Bestimmung der Signifikanzschwelle der LRT-Werte und Festlegung des Konfidenzintervalls des QTL

Die Bestimmung der Signifikanzschwelle erfolgte mit einem konservativen *P*-Wert von 0,001. In Folge wurden nur hochsignifikante LRT-Werte berücksichtigt und die Zahl falsch positiver Ergebnisse (Fehler 1. Art) auf ein Minimum begrenzt. Aufgrund der Verwendung von 1.143 Markerfenster musste der P-Wert zusätzlich noch mit der Bonferroni Korrektur für Mehrfachtests (DEWAN et al., 2007) modifiziert werden. Dies führte zu einem korrigierten *P*-Wert von $< 8{,}75^{-7}$ (d.h. 0,001/1.143). Für die Lokalisation des QTL wurden anschließend nur die LRT-Werte betrachtet, welche die berechnete LRT Signifikanzschwelle von 24,185 überschritten.

Um die Grenzen des Quantitativen Trait Locus an einem LRT-Maximum (LRT_{max}) festzustellen, wurde mit Hilfe des *logarithm of the odds* (LOD) Kriteriums das Konfidenzintervall (KI) eines QTL bestimmt (VAN OOIJEN, 1992). Die Grenzen des 2-LOD KI, welche mit einer Wahrscheinlichkeit von mehr als 95 % die gesuchte kausale Mutation enthalten, errechneten sich aus der Differenz zwischen dem LRT_{max} und 9,21. Ein LOD entsprach hierbei 4,605 (VISSCHER & GODDARD, 2004).

3.2.3.9. Identifikation eines kausalen Haplotypen

Zur Auswertung der cLDLA Daten wurde eine Tabelle angelehnt an MÜLLER et al. (2017) zu den einzelnen Individuen angelegt. Die folgenden Informationen wurden hierbei berücksichtigt: maternale und paternale Marker-Haplotypen des 40 SNP Fensters, das den signifikanten QTL beinhaltete, Diplotypeffektgröße (mittels Modells der Alternativhypothese und *ASREML* geschätzt) sowie den Haplo- und Diplotypen zugeordneten Indexwerte. Ergänzt wurden diese Parameter mit Werten für das Geschlecht, inklusive den zugehörigen Geschlechts-Effekten und den Informationen der polygenen Effekte. Die Tiere wurden zunächst anhand ihrer Diplotypeffekte in aufsteigender Reihenfolge sortiert. Anschließend erfolgte ein Vergleich der maternalen und paternalen Marker-Haplotypen der Tiere mit den niedrigsten Diplotypeffekt-Werten innerhalb des 40 SNP Fensters am Ziel-QTL. Anhand des gemeinsamen Haplotypenmusters und negativen Diplotypeffekten konnte der vermutlich kausale Haplotyp (im Folgenden Q) identifiziert werden. Infolgedessen wurde das Tierset analog zu MÜLLER et al. (2017) in drei Gruppen eingeteilt: die homozygote Gruppe Q/Q mit dem vermutlich kausalen Haplotypen am QTL, die heterozygote Gruppe q/Q mit jeweils einer Kopie des kausalen Haplotypen am QTL und die Gruppe q/q, die alle anderen Haplotypen widerspiegelte, die nichts mit dem kausalen Haplotypen gemeinsam hatte. Die 50K Marker-Haplotypen wurden anschließend von allen Tieren der Gruppen Q/Q und q/Q zusätzlich noch in der Region von 50–65 Mbp miteinander verglichen, um die exakten Grenzen des kausalen Haplotypen festzulegen.

3.2.4. Oxford Nanopore Technologie

Die Oxford Nanopore Technologie (ONT) zählt neben Pacific Biosciences (PacBio) und Single-molecule real-time sequencing (SMRT) zu den Third-Generation Sequenzierungsmethoden (siehe Kapitel 2.4). Die Long-Reads der Oxford Nanopore Technologie erreichen hierbei eine Länge von 500 bp bis zu 2,3 Mbp (PAYNE et al., 2018). Das Basiselement der ONT bildet eine sogenannte Flow-Cell. In der Flow-Cell befinden sich zwei Kammern, welche mit einer ionischen Lösung befüllt werden. Eine dazwischenliegende Membran separiert die beiden Kammern in eine *cis*- und eine *trans*-Seite, wobei die Kammer der *trans*-Seite immer eine positive Ladung gegenüber der *cis*-Seite aufweist (DEAMER et al., 2016). Perforiert wird die Membran mit einer unterschiedlich hohen Anzahl an biologischen Nanoporen, diese entstehen mit Hilfe porenbildender Proteine (z.B. α-Hämolysin) (OXFORD NANOPORE TECHNOLOGIES, 2020c). Diese natürlichen Proteine sind in einer Vielzahl von Organismen nachweisbar und ermöglichen den physiologischen Transfer von Molekülen in und aus der Zelle. Das Prinzip der ONT beruht auf dem passiven Molekültransport ausgelöst durch die unterschiedliche elektrische Ladung zwischen zwei durch eine Membran getrennten Räumen. In einem Organismus ist der Intrazellularraum aufgrund der hohen Chloridionen (Cl^-) Konzentration negativ geladen, während der Extrazellularbereich mit einem vermehrten Vorkommen von Natriumionen (Na^+) positiv ist. Der Stofftransport der Moleküle folgt bei elektrisch geladenen Teilchen neben den Gesetzen der Diffusion zusätzlich dem elektrischen Potenzialdifferenz über der Zellmembran (VON ENGELHARDT & BREVES, 2009). Der elektrische Ausgleichsstrom geladener Teilchen ermöglicht es, dass in der Flow-Cell die negativ geladene DNA oder RNA durch die Nanopore zur positiv geladenen *trans*-Seite strömt. Beim Durchtritt durch die Pore erzeugt jede der vier Nukleinbasen (Adenin, Thymin, Cytosin und Guanin) eine individuelle Spannungsveränderung, anhand derer im nächsten Schritt die jeweilige Basensequenz ermittelt wird (DEAMER et al., 2016).

Zur Durchführung der Sequenzierung wird das zuvor präparierte genetische Material (siehe Kapitel 3.2.4.1) auf die Flow-Cell aufgetragen. Ein an der Membran fixiertes Tether-Protein heftet sich an das DNA-Ende und führt diese an die Nanopore heran. Die Auftrennung des DNA-Doppelstrangs und die feste Bindung an die Nanopore erfolgt über das zuvor angehängte Motor-Protein (z.B. Polymerase oder Helikase). Beginnend mit einer angebrachten

Adapter-Sequenz startet nun der Übertritt des negativ geladenen DNA-Einzelstrangs von der negativ geladene cis- zur positiv geladenen trans-Seite. (Abbildung 2). Die gemessenen Spannungsveränderungen der einzelnen Nukleinsäuren werden anschließend durch das sogenannte Base-Calling (siehe Kapitel 3.2.4.2) in die jeweilige Base übersetzt. Die produzierten DNA Long-Reads können anschließend in weitere Analysen implementiert werden (DEAMER et al., 2016; OXFORD NANOPORE TECHNOLOGIES, 2020a).

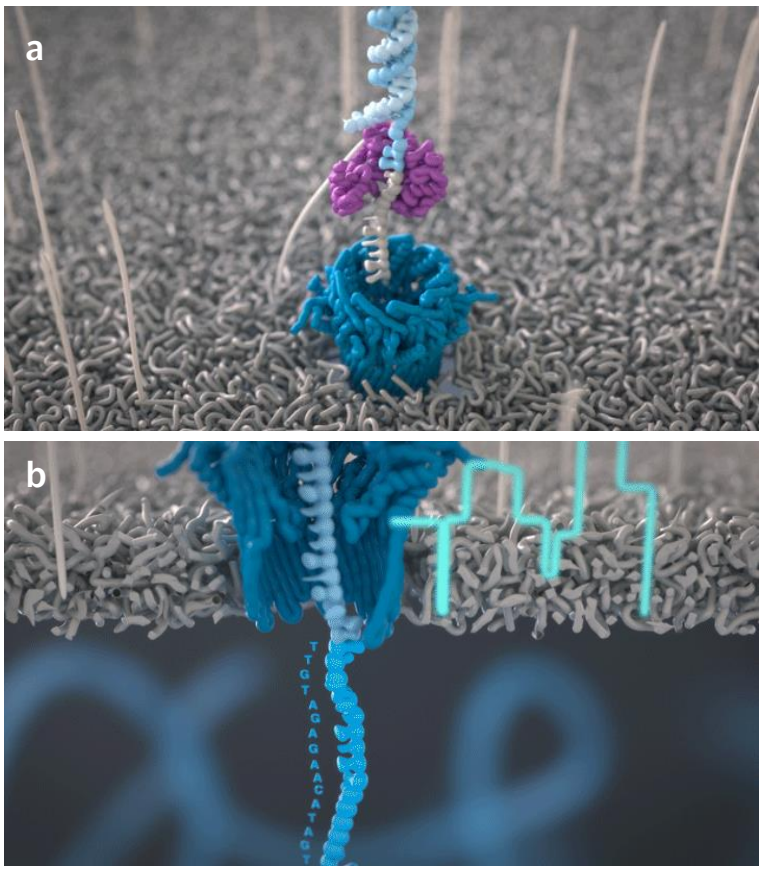


Abbildung 2: Funktionsweise der Oxford Nanopore Technologie
(Quelle:(OXFORD NANOPORE TECHNOLOGIES, 2020b))
a) Im oberen Bild ist die negative *cis*-Seite der Flow-Cell zu sehen. Der DNA (hellblau) wurde zunächst eine Adapter-Sequenz (grau) und ein Motorprotein (lila) angehängt. Das Tether-Protein der Membran (grauer Faden) führt das DNA-Gebilde an die Nanopore heran. Das Motorprotein trennt den DNA-Doppelstrang und bindet die DNA an die Nanopore. Nun erfolgt beginnend mit der Adapter-Sequenz der Durchtritt des negativ geladenen Einzelstranges zur positiven trans-Seite.
b) Im unteren Bild wird der erzeugte Spannungsunterschied je nach Base nochmals dargestellt (hellblaue Kurve). Dieser wird gemessen und mittels Base-Calling die eigentliche DNA-Sequenz generiert.

3.2.4.1. Präparation der DNA-Library

Nach Auswertung der Haplotypen wurden insgesamt vier Individuen des HF-Datensatzes mit archivierten Blut- und Spermaproben für eine Oxford Nanopore Sequenzierung ausgewählt. Aus der heterozygoten Gruppe q/Q

stammten die weiblichen Tiere DHFnano01 und DHFnano02. Aus der homozygoten Gruppe Q/Q wurde der Bulle DHFnano03 gewählt, als Vertreter des vermutlich kausalen Haplotypen. Zuletzt wurde das Tierset mit dem Bullen DHFnano04 vervollständigt, der die Gruppe q/q repräsentierte und dessen homozygoter Haplotyp nichts mit dem vermutlich kausalen Haplotypen gemeinsam hatte. Die Blut- und Spermaproben der Tiere wurden an das Laboratory of Functional Genome Analysis (LAFUGA) des LMU Gene Center München übergeben, um die Long-Read Oxford Nanopore Sequenzierung durchzuführen.

Zur Herstellung der DNA Library für die PromethION Flow-Cell wurde das Oxford Nanopore LSK109 Kit verwendet. Zuerst wurden das Motor-Protein und die Adapter-Sequenz an die DNA angehängt. Dieser Vorgang ist nur möglich, wenn das DNA-Ende einzelsträngig mit einer 3'-Orientierung und einem Poly-A-Schwanz vorliegt (ACHIMASTOU, 2019). Hierfür wurden im ersten Schritt ca. 3 µg der DNA-Probe mit dem Ultra II end-repaired Modul, zur Verfügung gestellt von New England Biolabs (Ipswich MA, USA), für zehn Minuten bei 20°C bearbeitet. Um weitere Reaktionen der Enzyme mit den nachfolgenden Reagenzien zu verhindern, wurden diese bei 65°C für 5 Minuten inaktiviert. Die Reinigung der DNA erfolgte durch Zugabe von 1 Volumen des AMPure XP Aufreinigungssystems (Beckman Coulter, Brea CA, USA), welches anschließend von der aufgereinigten und endreparierten DNA bei 55°C und 20 Minuten wieder gelöst wurde. Die Ligation des Motor-Proteins und der Adapter-Sequenz an das DNA-Ende wurde mit dem T4 Quick-Ligation Modul (New England Biolabs, Ipswich MA, USA) und Komponenten des LSK109 Sequencing Kit durchgeführt. Die finale Aufreinigung erfolgte durch erneute Zugabe von 0,45 Volumen des AMPure XP und der darauffolgenden Elution mit dem LSK109 EB Puffer für 20 Minuten bei 37°C. Circa 400 ng des fertigen DNA-Reaktionsgemisches wurden auf die PromethION Flow-Cell geladen und mit dem PromethION Beta Sequenzierer innerhalb von 72 Stunden vollständig sequenziert (ONT, Oxford, UK).

3.2.4.2. Base-Calling

Im letzten Schritt der DNA-Sequenzierung fand das sogenannte Base-Calling zur Ermittlung der Basensequenz eines jeden Tieres satt. Dieser Vorgang wurde mittels Graphical Processing Units (GPU) offline auf dem PromethION Server durchgeführt. Der verwendete Basecaller *GUPPY* Version 3.2.2 (PAYNE et al., 2020) übersetzte hierfür die Spannungswerte in die jeweilige Base Adenin (A), Thymin (T), Cytosin (C) oder Guanin (G). Am Ende des gesamten Sequenzierungsprozesses standen die ONT Long-Reads der vier Proben DHFnano01, DHFnano02, DHFnano03 und DHFnano04 für die weiterführenden Analysen als sogenannte FASTQ-Files zur Verfügung. Zusätzlich wurde aus den Sequenzdaten der beiden heterozygoten Kühe (DHFnano01 und DHFnano02) ein kombiniertes FASTQ-File (DHFnano01/02) mit dem Hintergrund erstellt, möglichst weite Bereiche des Genoms abzudecken.

3.2.5. Sequence-Reads Alignment (Mapping) zur Untersuchung des signifikanten Quantitativen Trait Locus

Ein wichtiger Punkt bei der Analyse von NGS-Daten ist das Sequenz Alignment (*Mapping*). Bei dieser Analyse werden Millionen sequenzierter DNA-Fragmente (*Reads*) innerhalb kurzer Zeit mit einer ausgewählten Referenzgenomsequenz abgeglichen. In der englischen Literatur wird dieser Vorgang als Read-Mapping (DNA-Fragment Kartierung) bezeichnet. Um Missverständnissen oder gar Verwechslungen zwischen der genetischen Kartierung (z.B. QTL-Kartierung) und der Sequenz-Read Kartierung vorzubeugen, wurde in diese Arbeit die Bezeichnung Reads-Alignment verwendet. Daher bezieht sich im Folgenden das Wort *Mapping* ausschließlich auf das Read-Mapping und das Wort *Kartierung* nur auf die genetische Kartierung.

Mit dem Ziel, kausale Mutationen zu identifizieren, die mit dem Merkmal paternaler Kalbeverlauf assoziiert werden könnten, erfolgte das Alignment sequenzierter Reads ausgewählter Proben gegen das *Bos taurus* Referenzgenom ARS-UCD1.2 (ROSEN et al., 2020). Das Datenset aus verschiedenen Rinderrassen beinhaltete sowohl ONT sequenzierte Long-Reads als auch paired Illumina sequenzierter Short-Reads. Die Nanopore Ganzgenomsequenzen stammten von den in dieser Arbeit sequenzierten Holsteinproben DHFnano01, DHFnano02, DHFnano03, DHFnano04 und DHFnano01/02, den internen ONT Proben der Rassen Kärntner Blondvieh (KBVnano05) sowie einer Fleckvieh Probe (FVnano06) publiziert aus Studie von GEHRKE et al. (2020) (siehe Tabelle 6). Um eine Verzerrung der Ergebnisse aufgrund der Sequenzierungstechnik ausschließen zu können, wurden zusätzlich paired Illumina Short-Read sequenzierte Genome in die Studie mitaufgenommen. Hierfür wurden die freiverfügbaren Ganzgenomsequenzen der in Kapitel 3.2.3.2 beschrieben 21 Holstein Proben (Anhang 1), einer Hereford Kuh (SRR8324584) und eines Brahman Zebu Bullen (SRR2016745) über das NCBI SRA heruntergeladen.

3.2.5.1. Alignment der Oxford Nanopore Long-Reads

Das *Alignment* wurde mit dem Programm *MINIMAP2* in der Version 2.17 (LI, 2018) gegen das ARS-UCD1.2 Assembly durchgeführt. Die entstandenen SAM (*Sequence Alignment Map*) Dateien wurden mit dem Programm *SAMTOOLS* in der Version 1.9 (LI et al., 2009) für die weiteren Schritte in das BAM (*Binary Alignment Map*) Format konvertiert. Das Programm *SAMTOOLS* wurde außerdem dazu verwendet, um eine Sortierung der BAM Dateien in die chromosomal richtige Reihenfolge vorzunehmen und eine statistische Auswertung durchzuführen. Im nächsten Schritt wurden, um falsche Ergebnisse zu vermeiden, die DNA-Enden um ein paar Basenpaare durch das sogenannte Trimming gekürzt. Dieser Schritt war notwendig, da das Base-Calling v.a. an den Enden der meisten DNA-Sequenzen falsche Basen erzeugt, z.B. durch sequenzierte Adapter-Sequenzen (CHOU & HOLMES, 2001; SCHUBERT et al., 2016). Die Durchführung des Trimmings erfolgte mit dem Programm *SICKLE* Version 1.33 (JOSHI & FASS, 2011). Zusätzlich wurden mögliche Artefakte des Mappings mit dem Programm *PICARD* Version 2.20.7 (BROAD INSTITUTE, 2019) entfernt. Das Programm *INTEGRATIVE GENOMICS VIEWER* (IGV) der Version 2.6.3 (THORVALDSDOTTIR et al., 2013) ermöglichte die graphische Darstellung der DNA-Fragmente und wurde zur Detektion von Mutationen herangezogen. Chromosomenabschnitte, die sich in IGV besonders prägnant darstellten, wurden anschließend in der *ENSEMBL* Datenbank des Genomic Variants Archiv (DGVa) überprüft. Der *ENSEMBL GENOME BROWSER* 99 (CUNNINGHAM et al., 2018) ist eine freiverfügbare Datenbank für verschiedenste Studien vertebraler Genome. Zur Spezifikation bestimmter Region im Genom kann auf verschiedene Datenbanken, z.B. zu Genen oder Strukturvarianten (DGVa), als auch auf Programme wie *BLAST*, *BLAT*, *VARIANT EFFECT PREDICTOR* (VEP) oder *BIOMART* zurückgegriffen werden.

3.2.5.2. Alignment der Illumina Short-Reads

Im Gegensatz zum Alignment der ONT-Daten wurden die Illumina sequenzierten Proben mit dem *BURROWS-WHEELER ALIGNER* Version 0.7.17 (LI & DURBIN, 2009) anstelle des Programms *MINIMAP2* gemappt. Als Referenzgenom wurde ebenfalls das Hereford Assembly ARS-UCD1.2 verwendet. Da die heruntergeladenen Sequenzen zunächst als sogenannte

SRA-Files vorlagen, erfolgte zunächst die Konvertierung in FASTQ Dateien mit dem Programm *FastQC* Version 0.11.8 (BABRAHAM INSTITUTE, 2017). Gleichzeitig wurde mittels *FastQC* eine erste Qualitätsbewertung der Sequenzen durchgeführt. Die darauffolgenden Arbeitsschritte entsprachen denen des Mappings der Nanopore Long-Reads und gliederten sich in Konvertierung der SAM zu BAM-Files, chromosomale Sortierung der BAM Dateien, Trimming der DNA-Enden und Entfernung von Artefakten. Verwendet wurden dieselben Programme wie in Kapitel 3.2.5.1. Die visuelle Untersuchung der gemappten Short-Reads wurde identisch zu den ONT Long-Reads in IGV (THORVALDSDOTTIR et al., 2013) durchgeführt. Auffällige Sequenzabschnitte wurden anschließend in der *Ensembl* DGVa überprüft (CUNNINGHAM et al., 2018).

3.2.5.3. Validierung des Alignments mit dem UOA_Angus_1 Assembly

Um sicherzustellen, dass keine negative Beeinflussung der Ergebnisse aufgrund des verwendete Hereford Referenzgenoms stattfand, wurde zusätzlich ein Alignment mit dem Angus Assembly UOA_Angus_1 durchgeführt. Das UOA_Anugs_1 wurde als weiteres Referenzgenom ausgewählt, da aufgrund der erreichten sehr hohen Contig N50 von 37,8 Mbp weitere Bereiche des Rindergenoms entschlüsselt werden konnten (LOW et al., 2020). Im Gegensatz dazu erreichte das Hereford ARS-UCD1.2 eine Contig N50 von 25,9 Mbp (ROSEN et al., 2020). Das Validierungs-Alignment wurde mit den ONT sequenzierten Genomen der beiden heterozygoten Kühe DHFnano01 und DHFnano02 gegen das Angus Assembly durchgeführt. Die Vorgehensweise und Anwendung der einzelnen Programme entsprach vollständig dem Alignment in Kapitel 3.2.5.1. Um die verschiedenen Mappinganalysen miteinander vergleichen zu können, wurden die Positionen auffälliger Strukturen im ARS-UCD1.2 auf das UOA_Angus_1 Assembly mit Hilfe des NCBI Genome Remapping Service konvertiert (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2018).

3.2.5.4. Analyse der Nanopore Sequenzen mittels *de novo* Assembly Techniken

In einem weiteren Analyseansatz wurden die ONT Ganzgenomsequenzen auf Basis von *de novo* Assembly Techniken weiter untersucht. Das verwendete Datenset enthielt sowohl hetero- (DHFnano01, DHFnano02 und DHFnano01/02) und homozygote (DHFnano03) Holstein Tiere mit Assoziation zu dem vermeintlich kausalen Haplotypen, als auch eine homozygote Probe (DHFnano04) ohne Verbindung zu dem kausalen Haplotypen. Zusätzlich wurden die beiden weiteren Rassen Kärntner Blondvieh (KBVnano04) und Fleckvieh (FVnano05) in das Probenset mitaufgenommen. Vor allem die Anzahl, N50 und Verteilung der Contigs pro Individuum wurden in diesem Verfahren genauer betrachtet.

Das *de novo* Assembling dient der Rekonstruktion der ursprünglichen Chromosomen- bzw. Genomsequenz einer Probe durch die Erstellung einer physikalischen Genomkarte. Um eine solche Karte zu erhalten, werden im ersten Schritt anhand sequenzierter Reads sogenannte Contigs produziert. Hierfür ermittelt der Assembler überlappende Abschnitte zwischen den DNA Reads und verbindet diese zu einem längeren Fragment, den Contigs (NATIONAL HUMAN GENOME RESEARCH INSTITUTE, 2011). Für die Assemblierung von Long-Reads bedeutet dies allerdings, dass der Abgleich der gesamten Sequenzen der einzelnen Reads nicht nur ausgesprochen zeit- und kostenintensiv wäre, sondern zudem große Teile der Speicherkapazität eines Rechensystems in Anspruch nehmen würde. Um daher eine leistungsfähige Analyse zu ermöglichen, unterteilen Assembler wie der *WTDBG2* (*READBEAN*) die Long-Reads in Gruppen (Bins) mit einer Länge von 256 bp (LI et al., 2009). Dies ermöglicht eine effiziente Speicherbelegung, da in 8 Bit $2^8 = 256$ Zeichen kodiert werden können. Acht Bit ergeben wiederrum ein Byte, wodurch pro Bin lediglich ein Byte der Speicherkapazität benötigt wird (PRECHT et al., 2004). Nun werden bei der Suche nach überlappenden Sequenzen leidglich die Bins miteinander verglichen und nicht die gesamte Sequenz der Reads. Die Erstellung der Contigs erfolgt anschließend durch Aneinanderreihung der Reads mit überlappenden Segmenten (RUAN & LI, 2019). Um nun eine möglichst durchgehende Sequenz eines Chromosoms bzw. Genoms zu erhalten, wird nun entweder die Position der Contigs mit Hilfe der Sequenz eines Referenzgenoms bestimmt, oder die Contigs erneut

anhand überlappender Sequenzen zu noch längeren Fragmenten (*Scaffolds*) verbunden (NG & KIRKNESS, 2010). In dieser Studie wurde die Position der Contigs durch Mapping an ein Referenzgenom bestimmt.

Ein wichtiges Qualitätskriterium der *de novo* Assemblierung ist die Contig N50. Zur Bestimmung der N50 werden die Contigs nach ihrer Länge in zwei Gruppen unterteilt, die jeweils 50 % des Genoms abdecken. Das heißt, in der ersten Gruppe befindet sich das längste, dann das zweitlängste Contig usw. bis 50% des Genoms bedeckt sind. Die N50 spiegelt nun die Länge des kürzesten Contigs der ersten Gruppe wider. Somit werden 50 % des Genoms von Contigs abgedeckt, die länger als das Contig sind, welches die N50 bestimmt. Daher steigt der Informationsgehalt eines Assemblies mit einer höheren N50, da in diesem Fall weniger unsequenzierte Lücken zwischen den durchschnittlich längeren Contigs enthalten sind (CASTRO & NG, 2017).

Die Contigs der insgesamt sieben Proben wurden jeweils in einem separaten Ansatz mit dem *WTDBG2* (READBEAN) Assembler der Version 2.5 (RUAN & LI, 2019) erstellt. Um sicherzustellen, dass der Assembler nicht durch angehängte Adaptersequenzen negativ beeinträchtigt wird, wurden diese vor dem eigentlichen Assembling entfernt. Das Programm *PORECHOP* Version 0.2.4 (WICK, 2018) mappte zu diesem Zweck einzelne Fragmente gegen bekannte Adaptersets. Im Falle eines komplementären Abschnitts wurden von diesem Fragment die Enden um ca. vier Basen getrimmt. Zusätzlich wurde ein Sicherheitskriterium von zwei Basen miteinkalkuliert, um eine vollständige Entfernung zu gewährleisten. Adaptersequenzen, die sich in der Mitte eines Reads befanden, wurden von dem Programm als Chimäre identifiziert und in zwei eigenständige Fragmente separiert. Waren die einzelnen Abschnitte zu kurz, wurden diese verworfen (WICK, 2018). Die Fehlerkorrektur der generierten Contigs wurde im Anschluss mit dem Programm *RACON* Version 1.3.3 (VASER et al., 2017) in drei Runden durchgeführt. *RACON* basiert auf den von LEE et al. (2002) und LEE (2003) bereits beschriebenen Partial Order Alignment (POA) Graphen. Zur Ausführung des Filterprozesses werden die gewählten Qualitätsparameter und sogenannte „query-to-target mappings" benötigt. Wobei „query" die Gesamtheit der Contigs darstellt und „target" jene Contigs, die im Zusammenhang mit der Fehlerkorrektur stehen.

Contigs mit hoher Fehlerrate wurden anhand des folgenden Algorithmus aussortiert:

$$|1 - \min(dq, dt)/\max(dq, dt)| > e$$

Wobei *dq* und *dt* jeweils die Länge der query oder target Contigs repräsentiert und *e* den spezifischen Grenzwert der Fehlerrate.

Anschließend wurden die gefilterten Contigs in nichtüberlappende Fenster entlang der Sequenz eingeteilt und der POA Graph konstruiert. Die finale Sequenz wurde durch Aneinanderreihung der einzelnen Fenster erstellt (VASER et al., 2017). Im letzten Schritt erfolgte die Zuweisung der Position eines jeden Contigs im Genom durch Alignment an das ARS-UCD1.2 Referenzgenom (ROSEN et al., 2020). Um auszuschließen, dass einzelne Contigs aufgrund des Hereford Referenzgenoms nicht gemappt werden konnten, wurde in einem zweiten Alignment ebenfalls das UOA_Angus_1 Assembly (LOW et al., 2020) als Referenzgenom eingesetzt.

3.2.6. Nachweis chromosomaler Aberrationen

Nach der Lokalisation des HF spezifischen signifikanten QTL und Identifikation eines kausalen Haplotypen assoziiert mit dem Merkmal paternaler Kalbeverlauf bestand der nächste Schritt darin, zugrundeliegende Mutationen zu detektieren. Die Analysen konzentrierten sich auf den Nachweis segmentaler Duplikationen (SD), struktureller Variationen (SV) sowie SNPs und einfachen Indels.

3.2.6.1. Identifikation segmentaler Duplikationen

Segmentale Duplikationen führen zu nicht-allelischen homologen Rekombinationsereignissen (NAHR) im Zuge des Crossing-overs. Die Folge sind Deletionen, Duplikationen und Inversionen in einem oder beiden Chromatiden oder in homologen Chromosomen. Darüber hinaus können dizentrische und azentrische Chromatide und Chromosomen entstehen, die eine letale Zygote nach sich ziehen (siehe Kapitel 2.5.1) (STANKIEWICZ & LUPSKI, 2002; FENG et al., 2017). Bisher veröffentliche Studien (LIU et al., 2009; WOMACK, 2012) zeigten eine hohe Anzahl segmentaler Duplikationen vor allem in der Region der Telomere des Chromosom 18 (*Bos taurus* UMD 3.1.1). Das Programm *SDDETECTOR* v0.2 (DALLERY et al., 2017) wurde verwendet, um segmentaler Duplikationen im ARS-UCD1.2 Assembly mit Fokus auf der Region des signifikanten QTL zu detektieren. Das Prinzip des *SDDETECTOR* beruht auf dem bioinformatischen Protokoll nach KHAJA et al. (2006). Im ersten Schritt erfolgte das Mapping des BTA18 gegen sich selbst unter Verwendung des Programms *MEGABLAST* (ZHANG et al., 2000). Das resultierende Output-File (XML) wurde anschließend in die Pipeline des *SDDETECTOR* zur Identifikation segmentaler Duplikationen implementiert. Nur jene Chromosomenabschnitte des BTA18, die bei der sogenannten Selbstkartierung eine Übereinstimmung ihrer Sequenzen von mindestens 90 % und eine Mindestlänge von 5 kb aufwiesen, erfüllten die Kriterien einer segmentalen Duplikation. Zur visuellen Veranschaulichung der segmentalen Duplikationen wurden diese im Anschluss mit Hilfe des Programms *CIRCOS* v0.52 (KRZYWINSKI et al., 2009) in einem zirkulären Graphen dargestellt.

3.2.6.2. Detektion struktureller Varianten

Wie in Kapitel 2.5 beschrieben, können bereits einfache Mutationen zu einer veränderten Proteintranskription der zugrundeliegenden Gene führen (GRAW, 2006). Um zusätzlich die Suche nach Kandidatengenen zu erleichtern, wurde die Region des kausalen Haplotypen auf SNPs untersucht, mit dem Ziel gekoppelte Gene aufzufinden (siehe Kapitel 2.3.2). Insgesamt wurden drei verschiedene Analyseansätze gewählt, um einerseits SNPs und andererseits strukturelle Varianten (SV) zu identifizieren: (1) Identifikation von SNPs und Indels mit dem GATK *HAPLOTYPECALLER* (HC), (2) Detektion von SV mit dem Programm *NANOVAR* in ONT sequenzierten Proben und (3) Detektion von SV unter Verwendung der Illumina sequenzierten HF Proben mit Hilfe der Programme *LUMPY*, *BREAKDANCER* und *MANTA*.

3.2.6.2.1. Identifikation von SNPs und Indels mit dem GATK *HaplotypeCaller*

Das Programm *HAPLOTYPECALLER* (HC) des GATK (MCKENNA et al., 2010) wurde verwendet, um Sequenzvarianten einer Probe im Vergleich zum Referenzgenom zu ermitteln. In dieser Analyse wurde das in Kapitel 3.2.3.2 generierte Tierset, bestehend aus 21 Illumina sequenzierten Holstein Proben, implementiert (Anhang 1) und mit der im selben Kapitel beschrieben Methode des *HAPLOTYPECALLER* ausgewertet. Jedoch ist zu beachten, dass in dieser Untersuchung die Option „includeNonVariantSites“ nicht verwendet wurde, da hier nur die Abweichungen vom Referenzgenom von Interesse waren. Zusätzlich wurde der Analysebereich auf 15 Mbp (55–60 Mbp) eingegrenzt.

Nach abgeschlossener Analyse wurden die detektierten Varianten der Illumina sequenzierten Proben mit Hilfe der ONT sequenzierten Proben in IGV visuell begutachtet. Es wurde versucht, nur die SNPs, die eine Assoziation zu dem Merkmal pKV aufwiesen, zu erhalten. Dabei wurde darauf geachtet, dass das Alternativallel in 100 % der Reads der homozygoten Probe DHFnano03 (Gruppe Q/Q) und in ca. 50 % der Reads der heterozygoten Proben DHFnano01 und DHFnano02 (Gruppe q/Q) nachweisbar war. In der Probe DHFnano04 aus der Gruppe q/q ohne Assoziation zum kausalen Haplotypen hingegen sollten alle Reads auf das Referenzallel hindeuten. Als weiteres Kriterium wurde überprüft, ob das Alternativallel auch in den ONT Sequenzen

der Rassen Kärntner Blondvieh und Fleckvieh aufzufinden war. Konnte das Alternativallel eines SNPs in einer dieser Rassen bestätigt werden, wurde dieser SNP ausgeschlossen, da eine HF spezifischen Assoziation somit widerlegt wurde. Um potenzielle Kandidatengene zu ermitteln, die in Kopplung zu den ausgewählten SNPs stehen, wurden die besten Kandidaten in der Bovine Genome Variation Database and Selective Signatures (BGVD) (CHEN et al., 2020) und der Gendatenbank des NCBI überprüft.

3.2.6.2.2. Nachweis struktureller Varianten in den ONT sequenzierten Holstein Daten

Die Kandidatenregion des Chromosom 18 wurde in den Proben DHFnano01, DHFnano02, DHFnano03 und DHFnano04 auf die Präsenz struktureller Varianten mit dem Programm *NANOVAR*-v1.3.8 (THAM et al., 2020) untersucht. Dieses besteht im Wesentlichen aus drei Arbeitsschritten: (1) dem Mapping von Long-Reads gegen ein ausgewähltes Referenzgenom, (2) der Detektion von SV durch Kalkulation der Read-Depth und (3) eine weitere Spezifikation und Validierung der identifizierten SV durch Anwendung des Artificial Neural Network (ANN) Algorithmus (THAM et al., 2020). Da das Mapping der Sequenzen in dieser Studie bereits mit *MINIMAP2* durchgeführt worden war, wurden die BAM Files direkt in den zweiten Schritt implementiert. *NANOVAR* untersuchte daraufhin die BAM Dateien auf unvollständige Mappingereignisse, diese äußern sich z.B. durch Reads, die getrennt an zwei Stellen im Genom (split Reads) oder unvollständig an einer Stelle im Genom (hard-clipped Reads) gemappt wurden (GATK-TEAM, 2019). Im Anschluss wurden auffällige Regionen anhand der folgenden sechs SV-Klassen charakterisiert: Deletion, Inversion, Tandem Duplikation, Insertion, Transposition und Translokation, und die Read-Depth in den chromosomalen Abschnitten mit atypischen Mappingereignissen berechnet (THAM et al., 2020). Die Read-Depth beschreibt die durchschnittliche Anzahl an Reads, die an einer bestimmten Position an das Referenzgenom gemappt werden konnten. In duplizierten Abschnitten ist, daher eine gesteigerte Read-Depth festzustellen, während in deletierten Regionen eine reduzierte Read-Depth auftritt (ESCARAMÍS et al., 2015). Zuletzt erfolgte die Validierung und Reduktion der falsch positiven SV mittels ANN-Algorithmus (THAM et al., 2020).

3.2.6.2.3. Nachweis struktureller Varianten der Illumina sequenzierten Holstein Daten

Nach Analyse der ONT sequenzierten Proben wurden anschließend die paired Illumina Ganzgenomsequenzen auf strukturelle Varianten hin untersucht. Hierfür wurden erneut die freiverfügbaren 21 HF Proben des NCBI SRA (Anhang 1) verwendet. Für diese Analyse kamen die drei Programme *LUMPY* v0.3.1 (LAYER et al., 2014) mit Anwendung der empfohlenen Skript-Pipeline *SMOOVE* (https://github.com/brentp/smoove), *BREAKDANCER* v1.4.5 (CHEN et al., 2009) und *MANTA* v1.6.0 (CHEN et al., 2016) zum Einsatz. Ähnlich zu dem Programm *NANOVAR* benötigen auch diese drei Programme zur Identifikation von Strukturvarianten die gemappte DNA in Form von BAM Dateien.

Die Detektion von SV mit Hilfe von *LUMPY* basierte auf einem Breakpoint Wahrscheinlichkeitsmodell. Dieses Modell bezog sowohl verschiedenste Signale des Mappings als auch Informationen bekannter Strukturvarianten in die Kalkulation neuer SV mit ein. Als Breakpoint wurden Basen bezeichnet, die nur im Genom der Probe in unmittelbarer Nachbarschaft zueinander standen, während sie im Referenzgenom weiter voneinander entfernt lagen (LAYER et al., 2014). Chromosomale Abschnitte mit strukturellen Varianten zeigen meist atypische Mappingsignale. Darunter Abweichungen der Orientierung, Anordnung und Länge von Paired-Reads, Trennung eines zusammenhängenden DNA-Fragments (Split-Reads) und gesteigerte (Duplikation) bzw. fehlende (Deletion) Read-Depth (ESCARAMÍS et al., 2015). Mit Hilfe von *LUMPY* wurden die zuvor kalkulierten Breakpoints auf die beschriebenen atypischen Mappingereignisse untersucht und auffällige Regionen zusätzlich mit bereits bekannten SV validiert (LAYER et al., 2014). Die Programme *BREAKDANCER* und *MANTA* funktionieren nach dem gleichen Prinzip wie *LUMPY*. Während bei der Anwendung der Pipeline *BREAKDANCER* vor allem kleine SV zwischen 10 und 100 bp detektiert werden können (CHEN et al., 2009), fokussiert sich *MANTA* auf die Identifikation weitreichender Varianten (CHEN et al., 2016).

4. Ergebnisse

4.1. Ergebnisse der cLDLA und Identifikation eines gemeinsamen kausalen Haplotypen

In der Region zwischen 54.233.746 bp bis 63.206.119 bp überschritten alle in der kombinierten Kopplungsungleichgewichts- und Kopplungsanalyse ermittelten LRT-Werte für das Merkmal paternaler Kalbeverlauf die Signifikanzschwelle von 24,185 (Abbildung 3). Der chromosomenweite *LRT*-Peak für dieses Merkmal an der Position 58.860.538 bp erzielte einen Wert von 122,9. Die per 2-LOD Kriterium ermittelten Grenzen des Konfidenzintervalls konnten an den Positionen 58.343.346 und 59.432.662 bp bestimmt werden und beinhalteten mit einer Sicherheit von 95 % den QTL (KI95%).

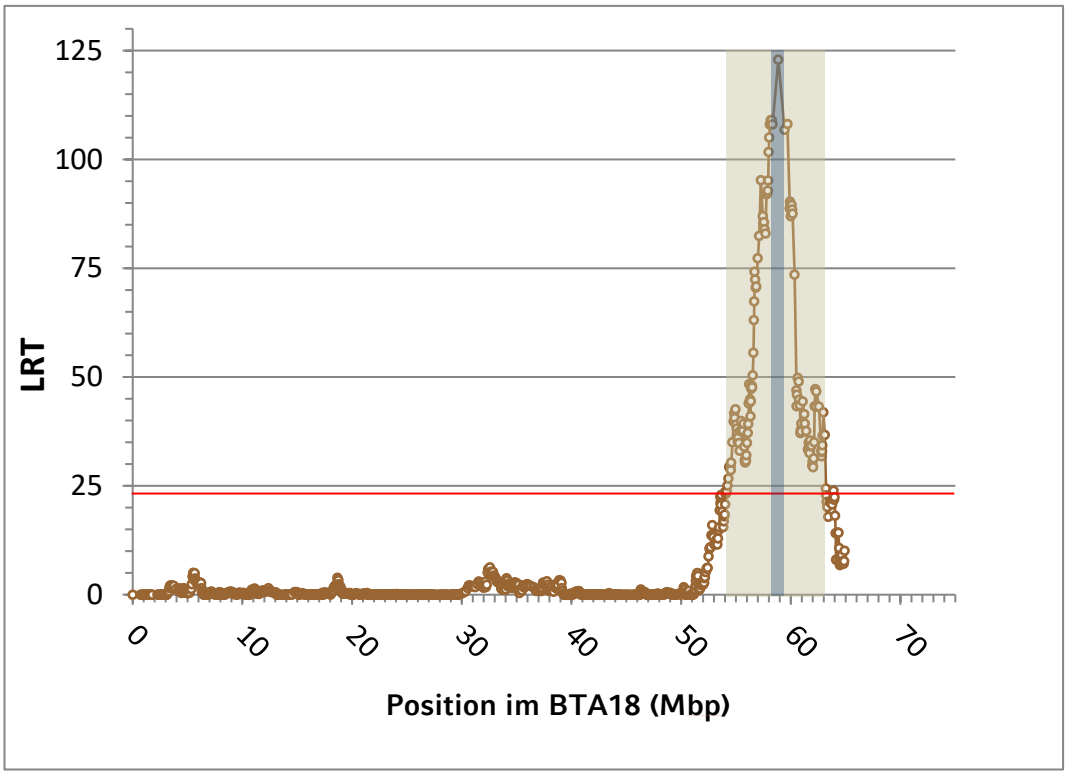


Abbildung 3: Verteilung der LRT-Werte für paternalen Kalbeverlauf auf Chromosom 18
Die X-Achse gibt die Position auf dem Chromosom 18 in Megabasen an, während die Y-Achse die LRT-Werte der cLDLA zeigt. Der LRT-Peak von 122,9 wurde an der Position 58.860.538 bp mit einem Konfidenzintervall von 58.343.346 – 59.432.662 bp lokalisiert (dunkelgrauer Balken). LRT-Werte, welche die Signifikanzschwelle von 24,185 (horizontale rote Linie) überschritten, befinden sich innerhalb des hellgrauen Balkens von 54.233.746 – 63.206.119 bp.

Zur Identifikation eines gemeinsamen kausalen Haplotypen wurde das 40 SNP Fenster, welches die Position des LRT_{max} als Mittelpunkt (d.h. zwischen Marker 20 und 21) beinhaltete, aller 2.697 Tiere miteinander verglichen. Anschließend wurde eine Sortierung des Datensets anhand der Diplotypeffekte vorgenommen. Der niedrigste Diplotypeffekt konnte bei -8,149 und der maximale bei +3,520 festgestellt werden. Insgesamt teilten 608 Tiere sowohl einen gemeinsamen Haplotyp als auch die niedrigsten Diplotypeffekt-Werte. Sechsundfünfzig dieser Tiere waren homozygot für den vermeintlich kausalen Haplotypen am QTL (d.h. Gruppe Q/Q) und erreichten Diplotypeffekte zwischen -8,149 bis -6,027. Die zweite Gruppe q/Q mit 552 Individuen trugen eine Kopie des kausalen Haplotypen und war folglich heterozygot am Ziel QTL mit Diplotypeffekten zwischen -5,661 und -1,766. Bei den restlichen 2.089 Tieren konnten zahlreiche unterschiedliche Haplotypen festgestellt werden mit Diplotypeffekt-Werten zwischen -1,766 und +3,520. Diese hatten allerdings nichts mit dem kausalen Haplotypen der vorherigen 608 Tiere gemeinsam und wurden folglich der Gruppe q/q zugeordnet.

Um die exakten Grenzen des vermutlich kausalen Haplotypen festzulegen, wurden die 50K Markerdaten der 608 Träger dieses Haplotypen (Q) (552 heterozygote und 56 homozygoten Individuen) in einem Bereich von 50–65 Mbp miteinander verglichen. Die Grenzen des überlappenden Chromosomenbereichs mit einem identischen Haplotypen in allen 608 Tieren konnten an den Positionen 57.922.208 bp und 60.057.741 bp festgelegt werden. Dieser chromosomale Abschnitt eines identischen Haplotypen von Tieren mit den niedrigsten Diplotypeffekten sowie maximalen LRT-Werten (incl. KI95%) wurde im Folgenden als „kausaler Haplotyp“ deklariert.

4.2. Ergebnisse der Sequenz-Mappinganalysen

4.2.1. Illumina Short-Read und Nanopore Long-Read Sequenzierung

Um die Region des kausalen Haplotypen weiter zu untersuchen, wurde zuerst die DNA ausgewählter Tiere mittels Oxford Nanopore Long-Read Technologie (ONT) sequenziert. Dieses Datenset beinhaltete sowohl zwei weibliche Individuen mit heterozygoten Haplotypen am Ziel QTL (Gruppe q/Q), als auch ein männliches homozygotes Tier aus der Gruppe Q/Q und einen männlichen Repräsentanten der Gruppe q/q. Zusätzlich wurde eine kombinierte Sequenz aus den Daten der beiden heterozygoten HF Kühe erstellt. Im Zuge der Sequenzierung konnten durchschnittlich 68,3 Gigabasen (Gbp) mit einer Reads N50 von 20.398 bp generiert werden. In der anschließenden Mappinganalyse konnten im Durchschnitt 88,59 % der DNA-Sequenzen mit einer mittleren Read-Coverage von 21,42 an das ARS-UCD1.2 gemappt werden. Eine detaillierte Qualitätsstatistik der einzelnen HF Proben und der ONT sequenzierten Proben der Vergleichsrassen Kärntner Blondvieh und Fleckvieh ist in Tabelle 6 aufgeführt.

Tabelle 6: Qualitätsparameter der nach der Oxford Nanopore Technologie sequenzierten Proben

Probe	Rasse	Geschlecht	Gbp [1]	Reads N50 [2]	Reads-Coverage [3]	Anzahl der gemappten Reads	Gemappte Reads in %
DHFnano01	Holstein	w	51,95	27.428	16,52	7.629.094	90,06
DHFnano02	Holstein	w	67,68	27.678	21,23	7.463.219	86,59
DHFnano03	Holstein	m	48,73	9.375	15,31	14.826.509	88,11
DHFnano04	Holstein	m	53,50	9.936	16,91	15.624.503	89,76
DHFnano01/02	Holstein	w	119,63	27.573	37,13	15.262.755	88,42
KBVnano05	Kärntner Blondvieh	m	71,61	11.816	17,78	26.582.401	85,46
FVnano06	Fleckvieh	m	53,64	12.100	12,83	11.660.445	91,41

[1] Gesamtgröße der generierten Genomsequenz in Gigabasen

[2] Zur Ermittlung der Reads N50 werden die Reads ähnlich der Contig N50 (siehe Kapitel 3.2.5.4) in zwei Gruppen eingeteilt. Die Reads N50 gibt nun die Länge des kürzesten Reads, aus der Gruppe der längsten Reads in Basenpaaren an, die benötigt wird, um 50% des Genoms abzudecken

[3] durchschnittliche Anzahl der Reads, die an einer bestimmten Position an das Referenzgenom mappen

Zusätzlich wurden vergleichend zu den ONT Sequenzen drei paired Illumina Short-Read Ganzgenomsequenzen in die Studie miteinbezogen. Die freiverfügbaren Daten stammten aus dem NCBI SRA und beinhalteten die Ganzgenomsequenzen der Rassen Holstein (ERR2694948), Hereford (SRR83245884) und Brahman Zebu (SRR2016745). Im Zuge des Mappings der Proben gegen das ARS-UCD1.2 wurden durchschnittlich 99,25 % der DNA Short-Reads erfolgreich gemappt mit einer mittleren Read-Coverage von 21,37. Eine Übersicht der Qualitätsstatistik dieser Proben ist in der nachstehenden Tabelle 7 zu finden.

Tabelle 7: Qualitätsparameter der heruntergeladenen paired Illumina Short-Read Sequenzen

Probe / Projektnr.	Rasse	Geschlecht	Gbp [1]	Reads-Coverage [2]	Anzahl der gemappten Reads	Gemappte Reads in %
ERR264948 / PRJEB27379	Holstein	m	151,69	45,20	865.469.672	99,56
SRR8324584 / PRJNA494431	Hereford	w	50,28	14,56	282.407.552	99,14
SRR2016745 / PRJNA277147	Brahman (Zebu)	m	14,67	4,37	125.663.248	99,05

[1] Gesamtgröße der generierten Genomsequenz in Gigabasen
[2] durchschnittliche Anzahl der Reads, die an einer bestimmten Position an das Referenzgenom mappen

4.2.2. Visuelle Untersuchung der Sequenz des Quantitativen Trait Locus

Im ersten Schritt wurde die umliegende Region des LRT-Peaks (58.860.538 bp) visuell begutachtet. Zu diesem Zweck wurden die gemappten Sequenzen aller Proben in das Programm *INTEGRATIVE GENOMICS VIEWER* (IGV) hochgeladen. Unabhängig von dem vorliegenden Haplotypen (homo- oder heterozygot) der HF Proben und der jeweiligen Rasse konnte in einem Bereich zwischen 58.846.130 bis 58.861.709 bp kein DNA-Fragment an die Referenz gemappt werden. Dieser chromosomale Abschnitt wurde im Folgenden als „16-Kb Lücke" deklariert. Die zuvor berechnete Position des LRT_{max} befand sich exakt innerhalb der 16-Kb Lücke, während das KI95% des QTL (58.343.346 - 59.432.662 bp) als auch die Grenzen des kausalen Haplotypen (57.922.208 – 60.057.741 bp) deutlich darüber hinaus reichten. Außerdem befand sich eine identische Lücke von 16.000 bp in den Illumina Short-Read

sequenzierten Proben verschiedener Rassen, wodurch ein Artefakt aufgrund der Sequenzierungsmethode ausgeschlossen werden konnte. In Abbildung 4 ist am Beispiel der Long-Read sequenzierten Proben DHFnano01 und KBVnano05 sowie der Illumina Short-Read Sequenzen der Proben ERR264948 (Holstein) und SRR8324584 (Hereford) die 16-Kb Lücke graphisch dargestellt. Die IGV Darstellung der 16-Kb Lücke aller Proben befindet sich zum Vergleich im Anhang 2 bis 4.

Bei genauerer Untersuchung der 16-Kb Lücke konnten zusätzlich in den beiden Holstein Proben DHFnano01 und DHFnano02 sowie in der des Kärntner Blondviehs zwei bis drei Reads pro Genom (d.h. jeweils 11 % der Reads pro Probe) mit sogenannten *linked supplementary Alignments* festgestellt werden. DNA-Fragmente die innerhalb komplexer Strukturen liegen, können im Zuge des Mappings durch Programme wie den *Burrows-Wheeler Aligner* oder MINIMAP2 fälschlicherweise separiert werden, wodurch die betroffenen Reads eine Lücke aufweisen. IGV ermöglicht es, derart betroffene Reads mit der Option „*linked supplementary Alignments*" darzustellen (PACIFIC BIOSCIENCES, 2017). Somit sind alle Reads, die *linked supplementary Alignments* aufwiesen, als kontinuierliche Reads ohne Lücke zu betrachten. Des Weiteren indizieren diese Reads, dass die 16-Kb Lücke aufgrund eines fehlerhaften Mappings der DNA Fragmente und nicht durch eine Deletion entstand.

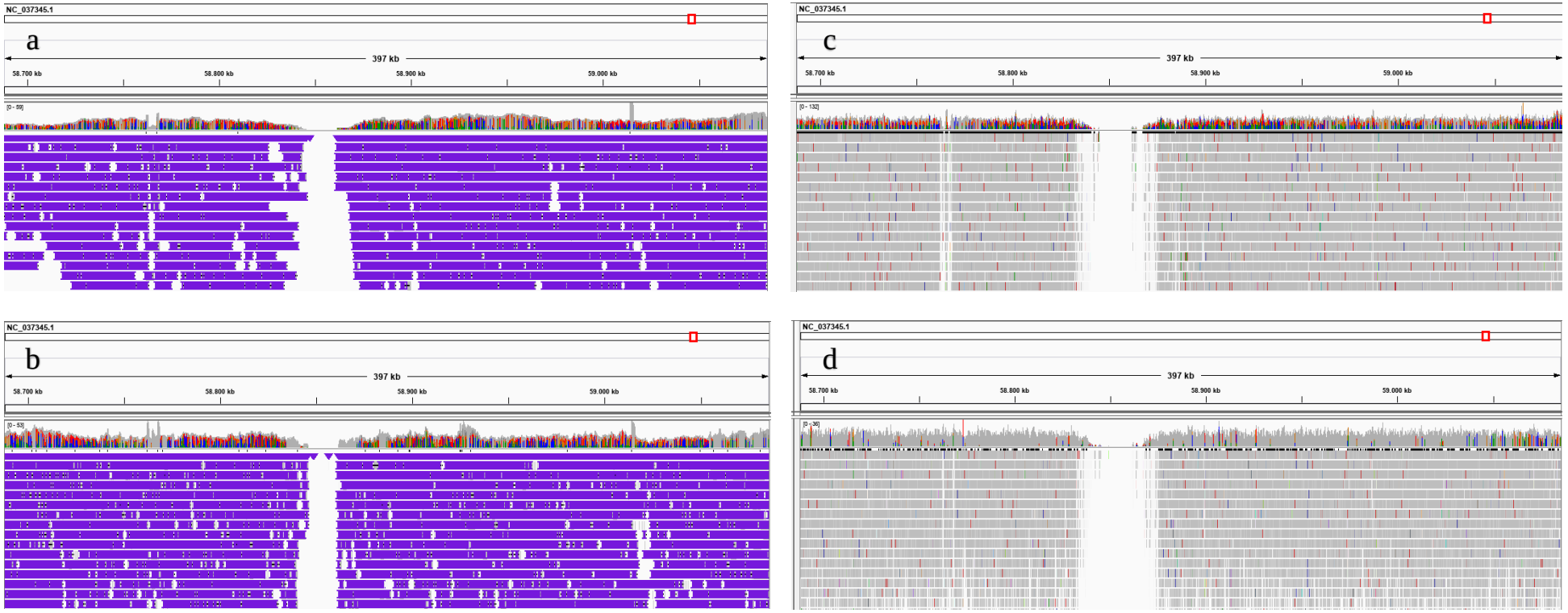


Abbildung 4: Graphische Darstellung der 16-Kb Lücke
Die Abbildungen zeigen jeweils den gleichen Bildausschnitt von 58.687.825 – 59.088.183 bp des Chromosom 18. Auf der linken Seite sind die ONT sequenzierten Proben des heterozygoten Tieres DHFnano01 (a) und des Kärntner Blondvieh (b), sowie auf der rechten Seite die Illumina Short-Read Sequenzen des Holstein Bullen ERR2694948 (c) und der Hereford Kuh SRR8324584 (d) zu sehen. In allen vier Abbildungen ist deutlich zu erkennen, dass innerhalb der 16-Kb Lücke (58.846.130 – 58.861.709 bp) kein DNA-Fragment an das ARS-UCD1.2 Referenzgenom gemappt wurde. Die Position des *LRT*-Peaks (58.860.538 bp) konnte exakt innerhalb dieser 16.000 bp lokalisiert werden, während das Konfidenzintervall (58.343.346 – 59.432.662 bp) darüber hinaus reicht.

4.2.3. Validierung der Ergebnisse mit dem UOA_Angus_1 Assembly

Mit der Publikation von LOW et al. (2020) stand zusätzlich ein weiteres *Bos taurus* Assembly (UOA_Angus_1) zur Verfügung, mit dem eine Validierung der bisher gewonnen Ergebnisse durchgeführt wurde. Für diese Analyse wurden die FASTQ Files der beiden heterozygoten Kühe ausgewählt und gegen das Angus Genom gemappt (siehe Kapitel 3.2.5.3). Jeweils 89,54 % (DHFnano01) bzw. 86,78 % (DHFnano02) der DNA Reads konnten mit einer Coverage von 17,34 bzw. 22,36 erfolgreich an die Referenz gemappt werden. Um vergleichbare Strukturen zu identifizieren, wurden die Positionen des ARS-UCD1.2 Assemblies mit Hilfe des NCBI *GENOME REMAPPING SERVICE* auf das Angus Referenzgenom konvertiert. Beim Vergleich der HD Markerpositionen auf dem ARS-UCD1.2 und dem UOA_Angus_1 konnte eine Sequenz in umgekehrter Reihenfolge beobachtet werden. Daher befand sich das KI95% des QTL zwischen 6.143.282 – 7.212.507 bp mit dem LRT_{max} an 6.697.993 bp. Die Start- und Endpositionen der 16-Kb Lücke wurden bei 6.496.595 bp bzw. 6.714.292 bp lokalisiert. Somit wurde die Region der 16-Kb Lücke im ARS-UCD1.2 Assembly (58.846.130 – 58.861.709 bp) bei Verwendung des Angus Referenzgenoms auf 217.697 bp ausgedehnt und mit langen Reads überdeckt. Obwohl innerhalb des Abschnitts der konvertierten ARS-UCD1.2 16-Kb Lücke Reads gemappt wurden, befand sich lediglich 2.517 bp vom distalen Rand der Lücke (Abbildung 5a und 5c rechts) entfernt eine vergleichbare Struktur ohne DNA-Fragmente (6.716.809 – 6.717.310 bp). Dieses Segment war mit 501 bp zwar wesentlich kleiner, befand sich jedoch innerhalb des KI95% (Abbildung 5b und 5d).

Bei genauerer Betrachtung der HD Markerpositionen und deren Verteilung auf dem Angusgenom, konnten 84 HD-Marker nicht eindeutig zugeordnet werden. Diese SNPs wiesen im Angus Assembly zwei Positionen auf, wohingegen dieselben 84 Marker eindeutig an eine einzige Position im ARS-UCD1.2 Assembly lokalisiert wurden. Für die weitere Analyse erfolgte zunächst die Untersuchung der 84 Marker im ARS-UCD1.2. Bei 76 der 84 SNPs konnte verteilt auf sieben Blöcke eine unmittelbare Nachbarschaft der Marker zueinander festgestellt werden. Anschließend wurde die Reihenfolge der 76 SNPs auf dem ARS-UCD1.2 Assembly mit der Reihenfolge derselben Marker auf dem Angusgenom verglichen. Hierbei stellte sich heraus, dass sich diese 76 Marker in derselben Reihenfolge befanden und auf dieselben sieben

Blöcke verteilten waren. Der einzige Unterschied bestand darin, dass die sieben Blöcke an zwei Stellen im Angusgenom auftraten. Darüber hinaus befanden sich drei Blöcke mit einer durchschnittlichen Länge von 24 kb innerhalb des KI95% des QTL (6.143.282 – 7.212.507 bp). Derart missverständliche Markerpositionen deuten darauf hin, dass der SNP bzw. die Region zwischen zwei mehrdeutigen Markern dupliziert wurde (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2020). Aufgrund der mehrdeutigen Markerpositionen konnte die Region des kausalen Haplotypen im Angusgenom nicht eindeutig lokalisiert werden, woraufhin auf weitere Analysen zur Identifikation kausaler Mutationen und/oder Kandidatengene mit diesem Assembly verzichtet wurde.

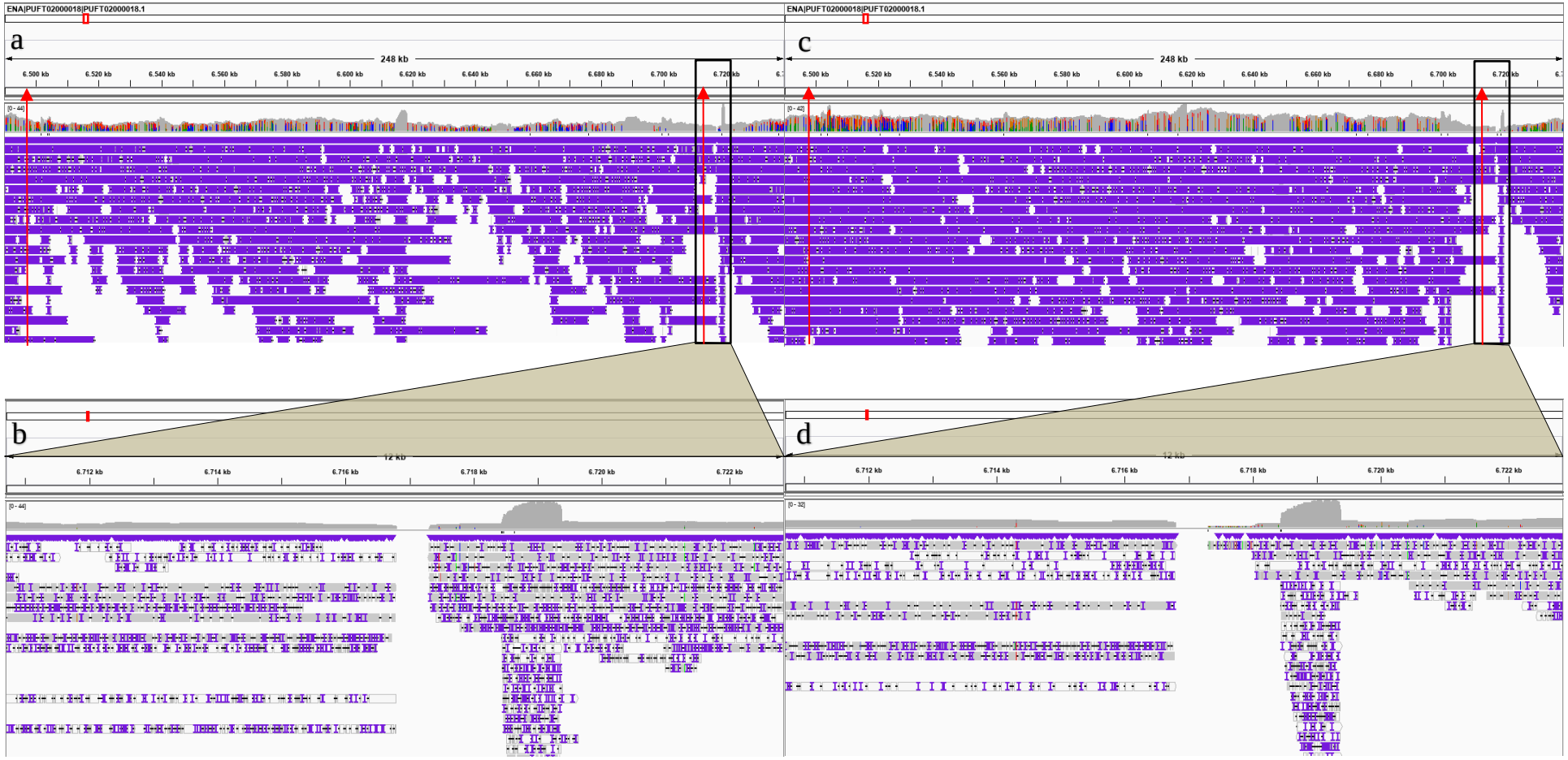


Abbildung 5: Ergebnisse des Sequenz Mappings an das *Bos taurus* UOA_Angus_1 Assembly
Die Abbildungen stellen das Read-Mapping der heterozygoten HF Kühe DHFnano01 (a und b) und DHFnano02 (c und d) dar. In a und c ist in der Region zwischen 6,49–6,74 Mbp die konvertierte Position der 16-Kb Lücke (6.496.595 – 6.714.292 bp) zu sehen (rote Pfeile). Der Bereich zwischen 6,71–6,72 Mbp wurde in b und d vergrößert dargestellt, um die Struktur zwischen 6.716.809 – 6.717.310 bp besser zu erkennen. Innerhalb dieser Region konnten auf 501 bp keine Reads lokalisiert werden. Dieser Abschnitt ohne Reads befand sich lediglich 2.517 bp von der 16-Kb Lücke entfernt.

4.3. Analyse des vermeintlich kausalen Haplotypen

Nach Abschluss der visuellen Untersuchung wurden verschiedene Analysen zur Aufklärung der zugrundeliegenden Struktur des vermutlich kausalen Haplotypen durchgeführt. Zuerst wurden die Abstände der einzelnen Markerpositionen auf dem Chromosom 18 ermittelt. Aufgrund der Tatsache, dass strukturelle Varianten und segmentale Duplikationen auch SNP Marker zusammen mit deren flankierenden Sequenzen duplizieren oder deletieren können, sind derart betroffene Marker für die Genotypisierung weniger nützlich und werden i.d.R. vom Markerset ausgeschlossen. Über das gesamte Chromosom konnte ein mittlerer Abstand zwischen benachbarten HD-Markern von 3.381 bp mit einer Standardabweichung von 4.098 bp beobachtet werden. Eine signifikant höhere Distanz ($P < 1\times10^{-4}$) konnte zwischen zwei benachbarten SNPs in der Region von 58,4 bis 59,9 Mbp festgestellt werden. Der mittlere Abstand lag hier bei 11.840 bp mit einer Standardabweichung von 12.213 bp. Diese 1,5 Mbp mit einer signifikant niedrigeren Markerdichte beinhalteten zu 94,8% das KI95% des QTL. Kein SNP konnte innerhalb der 16-Kb Lücke (58.846.130 – 58.861.709 bp) lokalisiert werden. Diese wurde lediglich durch die Marker BovineHD1800017266 an Position 58.836.794 bp und BovineHD1800017268 bei 58.877.026 bp mit einem Abstand von 40,2 kb flankiert. Abbildung 6 stellt den signifikanten Unterschied der Markerverteilung zwischen dem gesamten BTA18 und der abweichenden Region zwischen 58,4–59,9 Mbp dar.

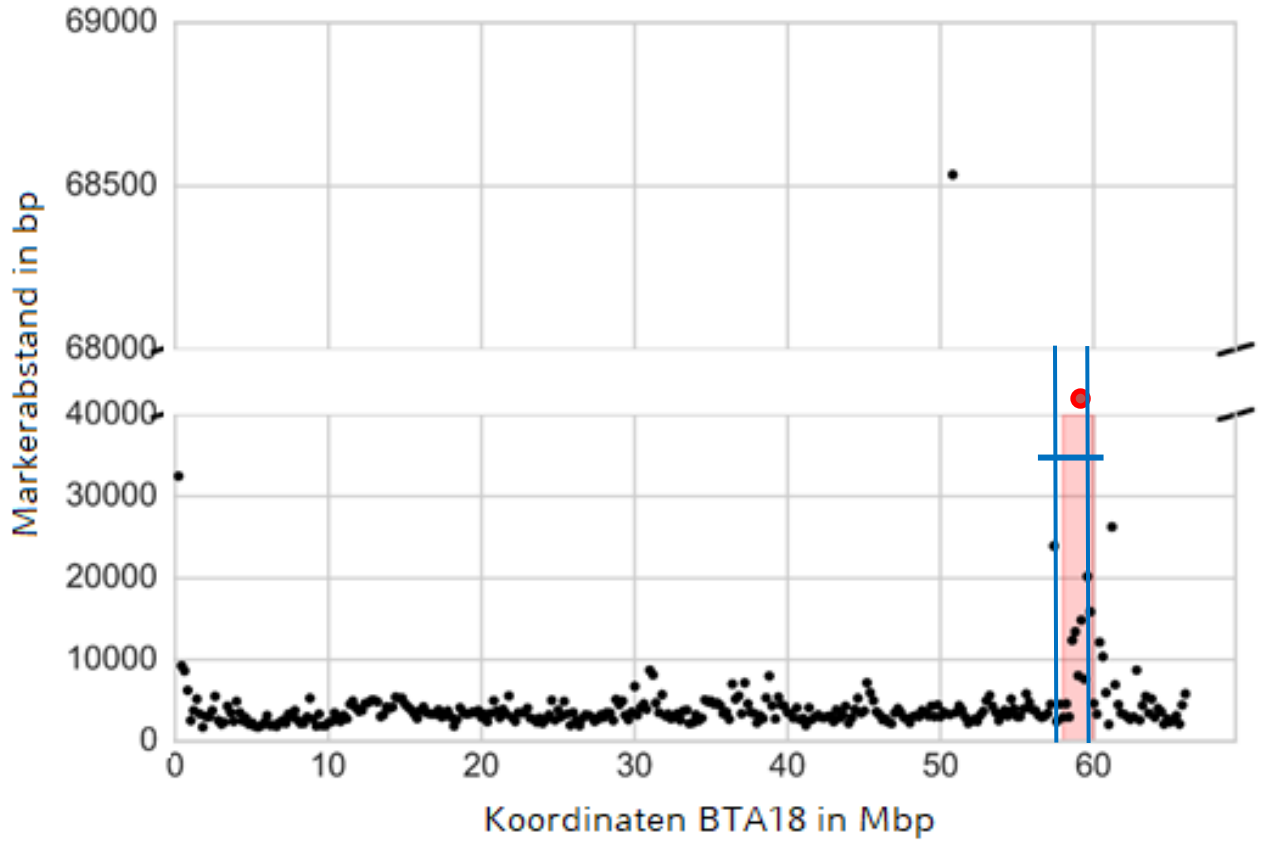


Abbildung 6: Punktdiagramm der Markerdichte auf Chromosom 18 im ARS-UCD1.2 Assembly
Die X-Achse des Diagramms zeigt die Position auf dem Chromosom 18 in Megabasen, während die Y-Achse den Markerabstand zwischen zwei Markern in Basenpaaren angibt. Die durchschnittliche Entfernung zwischen benachbarten SNPs auf dem gesamten Chromosom wurde bei 3.381 bp mit einer Standardabweichung von 4.908 bp gemessen. Der rote Balken zeigt die abweichende Region von 58,4 – 59,9 Mbp mit einer mittleren Markerdistanz von 11.840 bp und einer Standardabweichung von 12.213 bp. Der rote Punkt markiert den signifikanten Abstand von 40.200 bp zwischen den beiden Markern BovineHD1800017266 bei 58.836.794 bp und BovineHD1800017268 bei 58.877.026 bp, welche die 16-Kb Lücke flankieren. Die blauen vertikalen Linien spiegeln die Grenzen des KI95% des QTL (58.343.346 – 59.432.662 bp) wider, während die blaue horizontale Linie die Region des kausalen Haplotypen von 57.922.208 bis 60.057.741 bp zeigt.

Die sieben ONT sequenzierten Proben wurden in einem weiteren Schritt mit Hilfe von *de novo* Assembly Techniken analysiert. Im Gegensatz zu den in der Mappingstudie angewendeten Methoden wurden hierbei keine BAM Files erzeugt, sondern zunächst die DNA-Abschnitte der Sequenzierung zu längeren Contigs verbunden. Eine detaillierte Beschreibung der Contigbildung befindet sich in Kapitel 3.2.5.4. Die Proben erreichten eine durchschnittliche Contig N50 von 6.064.162 bp bei einer mittleren Gesamtsequenzlänge von 2,62 Gbp. Der statistische Bericht jeder einzelnen Probe ist in der nachstehenden Tabelle 8 aufgeführt.

Tabelle 8: Qualitätsstatistik der einzelnen Proben nach Durchführung der *de novo* Assembly Techniken

Probe	Rasse	Geschlecht	Gesamt-sequenzlänge (Gbp)	Contig N50 (Mbp)	Mittlere Contig Länge (bp)	Anzahl Contigs
DHFnano01	Holstein	w	2,68	4,56	480.238	5.494
DHFnano02	Holstein	w	2,65	10,28	543.098	4.883
DHFnano03	Holstein	m	2,55	1,83	373.119	6.836
DHFnano04	Holstein	m	2,57	2,43	460.898	5.578
DHFnano01/02	Holstein	w	2,68	16,95	434.302	6.181
KBVnano05	Kärntner Blondvieh	m	2,64	5,43	338.393	7.806
FVnano06	Fleckvieh	m	2,59	0,97	254.107	10.198

Die Contigs wurden anschließend gegen das ARS-UCD1.2 Genom gemappt und die Contiglänge sowie deren Verteilung isoliert für das Chromosom 18 genauer betrachtet. Wie bereits in den bisherigen Analysen hoben sich die Region zwischen 58–60 Mbp deutlich vom restlichen Chromosom ab. Innerhalb dieser 2 Mbp erreichte kein Contig eine Länge von mehr als 10 kb trotz einer hohen mittleren Contig N50 von mehr als 6 Mbp (Abbildung 7). Somit blieb die Contiglänge innerhalb der kritischen 2 Mbp, welche einerseits das KI95% des QTL vollständig beinhaltete und andererseits nahezu die gesamte Region des kausalen Haplotypen abdeckte, kleiner als die durchschnittliche genomweite Reads N50 der ONT sequenzierten Long-Reads (siehe Tabelle 6). Dieser Sequenzabschnitt über 2 Mbp trat vergleichbar zu der

16-Kb Lücke unabhängig von den Haplotypen der Holstein Individuen und von der jeweiligen Rasse der Proben auf. Um Abweichungen aufgrund des verwendeten ARS-UCD1.2 Genoms zu überprüfen, wurden die Contigs der sieben Proben zusätzlich gegen das UOA_Angus_1 Assembly gemappt. Nichtsdestotrotz befand sich unter Berücksichtigung der umgekehrten Basenabfolge bei Verwendung des Angusgenoms an derselben Position auf 2 Mbp Länge (4,5 – 6,5 Mbp) derselbe Abschnitt mit kurzen Contigs (Abbildung 7).

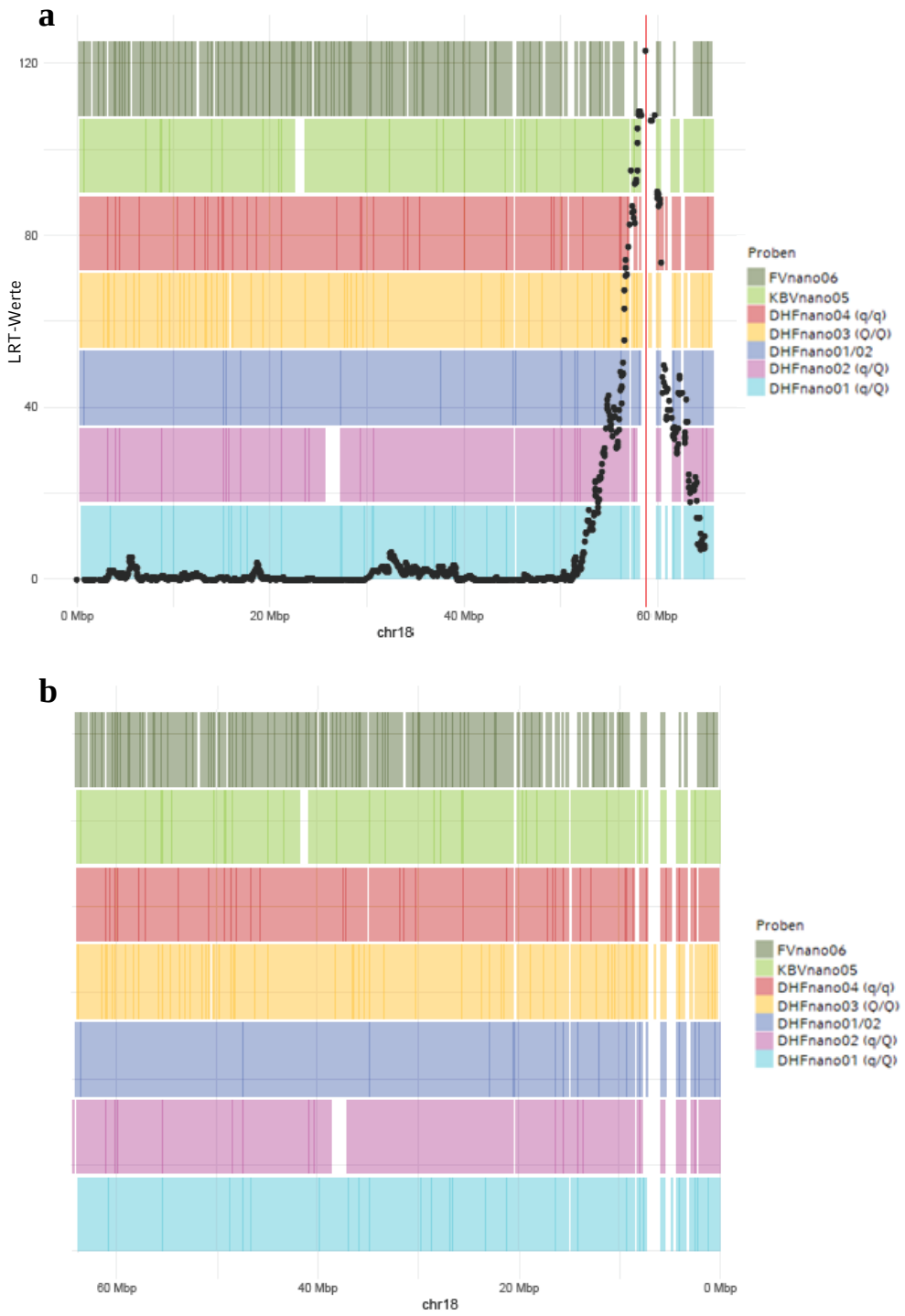


Abbildung 7: Grafische Darstellung der Contigverteilung auf Chromosom 18
Die ONT sequenzierten Proben von Holstein, Kärntner Blondvieh und Fleckvieh wurden einerseits gegen das ARS-UCD1.2 (a) und andererseits gegen das UOA_Angus_1 Assembly (b) gemappt. Die X-Achse zeigt die Position auf Chromosom 18 in Megabasen. Zusätzlich sind in (a) auf der Y-Achse die LRT-Werte der cLDLA angegeben. Der *LRT*-Peak an 58.860.538 bp ist anhand der roten vertikalen Linie dargestellt. Weiße Balken markieren jene Bereiche im Chromosom, in denen kein Contig eine Länge von 10 kb überschritt und folglich keine kontinuierliche Sequenz gebildet werden konnte. Farbintensivere Balken weisen auf eine erhöhte Anzahl kürzerer Contigs hin, die jedoch länger als 10 kb sind.

4.4. Identifikation von chromosomalen Aberrationen, SNPs und Indels auf Chromosom 18

Strukturelle Chromosomenaberrationen entstehen durch fehlerhafte Reparationsversuchen von einem oder mehreren DNA Doppelstrangbrüchen. In Folge kann eine Umstrukturierung der Genanordnung sowie der Basensequenz auftreten, wodurch die Expression der zugrundeliegend Gene mehr oder weniger stark beeinflusst wird (siehe Kapitel 2.5) (GRAW, 2006). Aus diesem Grund wurde die Region des kausalen Haplotypen auf die Präsenz chromosomaler Aberrationen wie segmentaler Duplikationen (SD) und einfacher struktureller Varianten (SV) untersucht, um Chromosomenabweichungen zu identifizieren, die möglicherweise mit dem Merkmal paternaler Kalbeverlauf in HF Rindern assoziiert sind. Die hierfür verwendeten Analysemethoden können in den Kapiteln 3.2.6.1, 3.2.6.2.2 sowie 3.2.6.2.3 im Detail nachgelesen werden. Zusätzlich wurden die ONT sequenzierten Holsteinproben nach dem in Kapitel 3.2.6.2.1 erläuterten Verfahren nach möglichen kausalen SNPs und einfachen Indels in der kritischen Region von 55–60 bp untersucht.

4.4.1. Detektion segmentaler Duplikationen

Segmentale Duplikationen (SD) sind eine Form der komplexen Mutationen, die unter anderem zu atypischen Crossing-over Vorgängen (z.B. nicht-allelische homologe Rekombination – NAHR) zwischen homologen Abschnitten führen können. In Folge der NAHR kann unter anderem eine veränderte Expression der betroffenen Gene beobachtet werden (siehe Kapitel 2.5.1) (STANKIEWICZ & LUPSKI, 2002). Vor diesem Hintergrund wurde ein negativer Einfluss segmentaler Duplikationen auf den Kalbeverlauf in Holstein-Friesian Rindern in Betracht gezogen. Der Nachweis der SDs erfolgte durch Mapping des BTA18 gegen sich selbst mit Hilfe der Programme *MEGABLAST* und *SDDETECTOR* nach der im Kapitel 3.2.6.1 beschriebenen Vorgehensweise. Am distalen Chromosomenende und in einem 2 Mbp langen Abschnitt zwischen 58 und 60 Mbp konnte eine signifikant höhere Anzahl ($P < 1\times10^{-4}$) segmentaler Duplikationen im Vergleich zum gesamten Chromosom detektiert werden (Abbildung 8b). Die jeweiligen Positionen der duplizierten Sequenzabschnitte einer SD sind in Form eines Bogens in der

zirkulären Abbildung 8a zu sehen. Hier geht klar hervor, dass das KI95% des QTL eine Vielzahl segmentaler Duplikationen beinhaltete (rote Bögen in Abbildung 8a). Darüber hinaus deckte sich der SD reiche Abschnitt zwischen 58 und 60 Mbp mit der kritischen Region einer signifikant geringeren Markerdichte und dem vermehrten Auftreten kurzer Contigs.

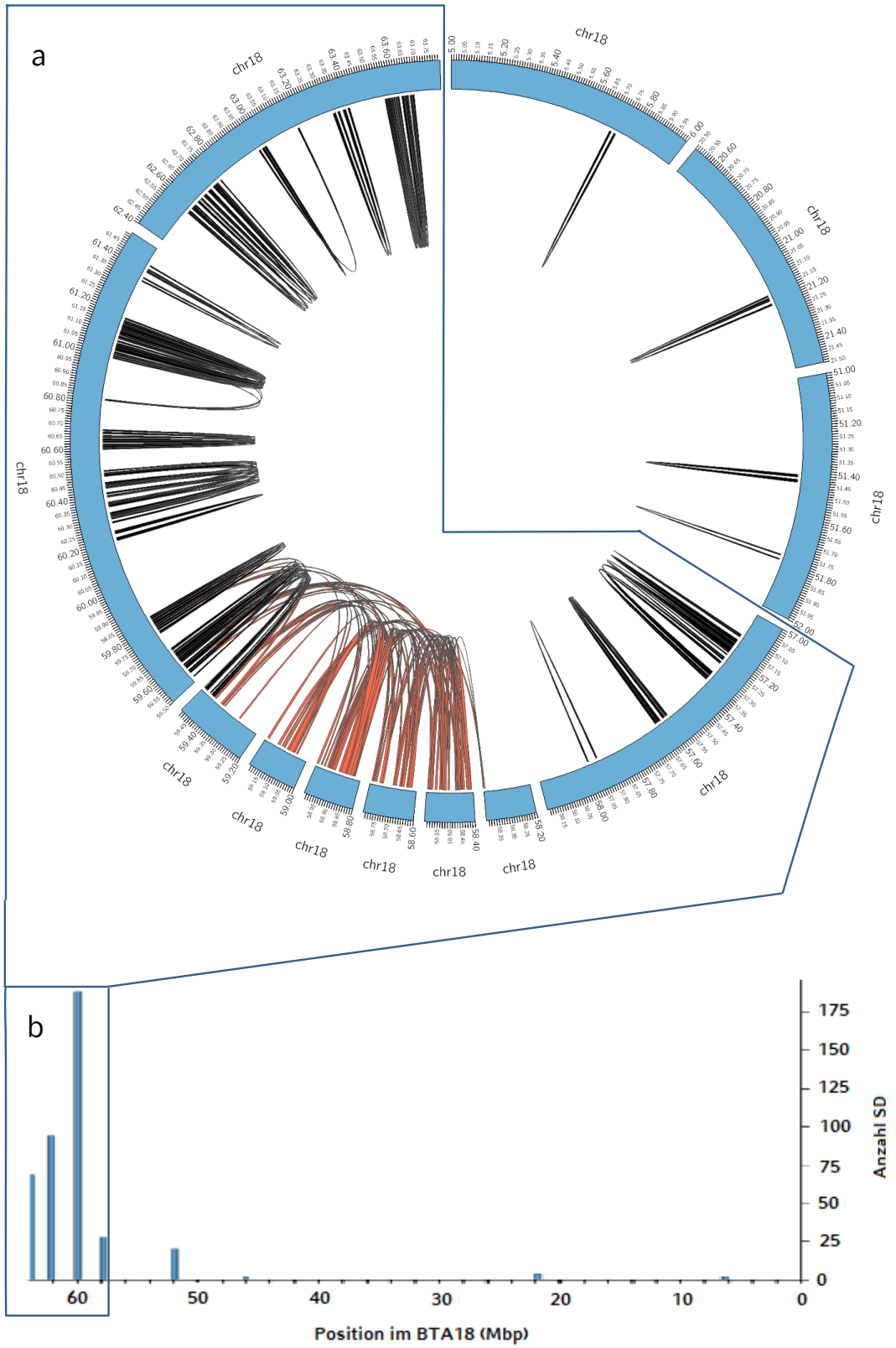


Abbildung 8: Verteilung segmentaler Duplikationen auf Chromosom 18
a) Zirkuläre Darstellung des BTA18. Zu beachten ist, dass chromosomale Regionen die keine SD aufwiesen, in der zirkulären Abbildung nicht dargestellt werden. Die Bögen innerhalb des zirkulären Chromosoms entsprechen den identifizierten segmentalen Duplikationen. Die Start- und Endpunkte eines Bogens markieren die Positionen der duplizierten Abschnitte. Eine erhöhte Anzahl konnte sowohl am distalen Chromosomenende als auch zwischen 58 und 60 Mbp festgestellt werden. SD im KI95% des QTL (58.343.346 – 59.432.662 bp) sind rot markiert.
b) Zur Verdeutlichung der relativen Verteilung der SD über das gesamte Chromosom 18 ist die Anzahl der segmentalen Duplikationen nochmals in einem Diagramm dargestellt (X-Achse Position auf dem Chromosom in Megabasen, Y-Achse Anzahl der SD). Dabei wurde jede SD doppelt gezählt, um sowohl die Donor- als auch die Akzeptor-Sequenz zu berücksichtigen.

4.4.2. Identifikation potenzieller Kandidaten-SNPs

Mit Hilfe des GATK *HAPLOTYPECALLER* konnten insgesamt 155 SNPs und Indels in einer Region von 55–60 Mbp lokalisiert werden. Die Überprüfung der HF spezifischen Assoziation zu dem Merkmal paternaler Kalbeverlauf erfolgte durch die visuelle Begutachtung der Allele in den ONT sequenzierten Proben mit Hilfe des Programms IGV. Insgesamt erfüllten vier SNPs (Tabelle 9) die festgelegten Kriterien, d.h. das Alternativallel eines SNPs lag homozygot in der Probe DHFnano03 vor bzw. heterozygot in den Proben DHFnano01 und DHFnano02, während die Proben DHFnano04, KBVnano05 und FVnano06 das Alternativallel nicht aufwiesen. Bei Überprüfung der SNPs auf potenzielle Kandidatengene mit Hilfe der Bovinen Genome Variation Database (BGVD) und der Gendatenbank des NCBI konnten die folgenden Gene ermittelt werden: In unmittelbarer Nähe zum ersten SNP rs381577268 (57.816.713 bp) befand sich das *Zink Finger Protein 613* (*ZNF613*) Gen zwischen 57.774.874 und 57816078 bp. Das *Zink Finger Protein 665-like* (*ZNF665-like*) Gen zwischen 59.513.790 und 59.550.728 bp konnte sowohl in der Nähe des SNP rs464221818 bei 59.329.176 bp als auch in der des SNP rs472502785 bei 59.345.689 bp lokalisiert werden. In der Nähe der Position des SNP rs381878735 (59.574.329 bp) konnten gleich zwei Gene detektiert werden, zum einen das *Zink Finger Protein 665 – ZNF665* Gen (59.513.790 – 59.550.728 bp) und das *Zink Finger Protein 677-like – ZNF677-like* Gen (59.416.313 – 59.431.145 bp). In der folgenden Tabelle 9 sind die vier potenziellen Kandidaten-SNPs im Detail aufgeführt.

Tabelle 9: Auflistung der Kandidaten-SNPs in der Region des kausalen Haplotypen

SNP Position (bp)	rs-Nummer	MAF [1]	Referenzallel	Alternativallel	Gen
57.816.137	rs381577268	0,010	C	T	*ZNF613*
59.329.176	rs464221818	0,010	C	T	*ZNF665-like*
59.345.689	rs472502785	0,016	T	C	*ZNF665-like*
59.574.329	rs381878735	0,011	A	T	*ZNF665, ZNF677-like*

[1] Minor Allele Frequency auf Datenbasis der Bovine Genome Variation Database and Selective Signatures (BGVD) (http://animal.nwsuaf.edu.cn/code/index.php/BosVar)

4.4.3. Strukturelle Varianten in den ONT sequenzierten Holstein Proben

Mit dem Ziel, strukturelle Varianten zu identifizieren, wurden die DNA Long-Reads der HF Proben mit dem Programm *NANOVAR* analysiert. Insgesamt konnten in allen Proben 33 SV detektiert werden. Diese verteilten sich wie folgt auf die einzelnen Proben: zehn (DHFnano01) bzw. neun (DHFnano02) SV wurden in den Sequenzen der heterozygoten Kühe nachgewiesen, sowie jeweils sieben SV in den homozygoten Bullen DHFnano03 (Gruppe Q/Q) und DHFnano04 (Gruppe q/q). Eine Auflistung der einzelnen SV pro Individuum befindet sich in der nachstehenden Tabelle 10. Obwohl mit keiner dieser Strukturvarianten die vollständige Übereinstimmung mit der erwarteten QTL-Allelverteilung gelang, konnte in drei Individuen eine gemeinsame Deletion lokalisiert werden. Diese Tiere waren entweder Träger einer Kopie des kausalen Haplotypen, d.h. heterozygot (DHFnano01 und DHFnano02) oder homozygot am Ziel QTL (DHFnano03). Nichtsdestotrotz musste diese Deletion als Kandidatenmutation ausgeschlossen werden, da sie in allen drei Proben homozygot vorlag. Im Falle einer kausalen Mutation des Kalbekomplexes wäre zu erwarten, dass die Deletion in den beiden heterozygoten Kühen nur in einem der beiden Haplotypen nachweisbar wäre.

Tabelle 10: Identifizierte strukturelle Varianten auf dem Chromosom 18 in den Holsteinproben DHFnano01, DHFnano02, DHFnano03 und DHFnano04
Die heterozygoten Proben DHFnano01 und DHFnano02 sind Träger von nur einer Kopie des kausalen Haplotypen (Gruppe q/Q), die homozygote Probe DHFnano03 stammt aus der Gruppe Q/Q mit Assoziation zu dem Merkmal paternaler Kalbeverlauf, während der Haplotyp der homozygoten Probe DHFnano04 der Gruppe q/q nichts mit dem kausalen Haplotypen gemeinsam hat.

SV [1]	Startposition (bp)	Endposition (bp)	Genotyp [2] DHFnano01	Genotyp DHFnano02	Genotyp DHFnano03	Genotyp DHFnano04
INV	27.628.586	58.763.694	0/1	0/0	0/0	0/0
INV	35.257.761	59.228.154	0/0	0/1	0/0	0/0
DEL*	58.484.966	58.485.244	1/1	1/1	1/1	0/0
DEL	58.762.192	58.767.668	0/0	0/0	0/1	0/0
DEL	58.762.211	58.768.920	0/0	0/1	0/0	0/0
DEL	58.762.211	58.768.937	0/1	0/0	0/0	0/0
DEL	58.812.976	58.813.026	0/0	1/1	0/0	0/0
DEL	58.812.981	58.813.024	0/0	0/0	0/0	1/1
DEL	58.812.982	58.813.027	0/0	0/0	1/1	0/0
DEL	58.812.983	58.813.027	1/1	0/0	0/0	0/0
DUP	58.889.079	58.889.174	0/1	0/0	0/0	0/0
DUP	58.889.080	58.889.180	0/0	0/0	0/0	0/1
DUP	58.889.082	58.889.171	0/0	0/1	0/0	0/0
DUP	58.890.160	58.890.282	0/0	0/0	0/0	1/1
DUP	58.890.167	58.890.280	1/1	0/0	0/0	0/0
DUP	58.890.170	58.890.352	0/0	0/0	1/1	0/0
DUP	58.890.174	58.890.353	0/0	1/1	0/0	0/0
DEL	58.910.339	58.910.384	0/0	0/0	0/0	1/1
DEL	58.910.347	58.910.391	0/0	1/1	0/0	0/0

DEL	58.910.348	58.910.391	1/1	0/0	0/0	0/0
DEL	58.910.353	58.910.401	0/0	0/0	1/1	0/0
DEL	59.123.315	61.313.922	0/0	0/0	0/0	0/1
DUP	59.219.123	59.219.148	0/1	0/0	0/0	0/0
DEL	59.226.555	59.226.595	0/0	0/1	0/0	0/0
DEL	59.226.562	59.226.609	0/0	0/0	0/0	1/1
DEL	59.226.567	59.226.603	0/1	0/0	0/0	0/0
DEL	59.409.006	59.409.112	0/1	0/0	0/0	0/0
DEL	59.409.009	59.409.114	0/0	0/1	0/0	0/0
DEL	59.414.242	59.414.415	0/0	0/0	0/1	0/0
DEL	59.414.252	59.415.012	0/0	0/0	0/0	0/1
DEL	59.414.878	59.415.004	0/0	0/0	0/1	0/0

[1] INV = Inversion; DEL = Deletion; DUP = Duplikation
[2] 0/0 = Alternativallel wurde nicht nachgewiesen; 0/1 = Alternativallel liegt heterozygot vor; 1/1 = Alternativallel liegt homozygot vor
* gemeinsame DEL in Trägern des kausalen Haplotypen

4.4.4. Detektion struktureller Varianten in den Illumina sequenzierten Holstein Proben

Neben den ONT sequenzierten Holstein Proben wurden zusätzlich paired Illumina Ganzgenomsequenzen auf strukturelle Varianten untersucht. Hierfür wurden die in Kapitel 3.2.3.2 beschriebenen 21 HF Proben des NCBI SRA mit Hilfe der Programme LUMPY, BREAKDANCER und MANTA analysiert (siehe Kapitel 3.2.6.2.2). Obwohl beim Vergleich der Marker-Haplotypen zwei (ERR2694951 und SRR4449830) der 21 Tiere den heterozygoten kausalen Haplotypen aufwiesen, konnte in keiner der Proben eine strukturelle Variante nachgewiesen werden.

5. Diskussion

In den letzten 20 Jahren wurden in einer Vielzahl von Studien die genetischen Hintergründe des Kalbekomplexes in Holstein-Friesian Rindern untersucht. Hierbei erwies sich vor allem das Chromosom 18 als besonders prägnant. COLE et al. (2009) gelang es, auf diesem Chromosom einen signifikanten QTL, assoziiert mit Kalbeverlaufs- und Körperkonstitutionsmerkmalen, zu identifizieren. Mehrere Studien konnten in den folgenden Jahren diese Entdeckung bestätigen (MAO et al., 2016; MÜLLER et al., 2017; PURFIELD et al., 2020) und den signifikanten QTL auf einen Bereich zwischen 50 und 60 Mbp lokalisieren. Darüber hinaus wurde nach unserem Kenntnisstand dieser QTL bisher ausschließlich in reinrassigen Holsteinrindern und HF Veredelungskreuzungen nachgewiesen. Jedoch wurden bis zum jetzigen Zeitpunkt weder Kandidatengene noch kausale Mutationen publiziert, die diesen QTL eindeutig entschlüsseln konnten. Ziel dieser Studie war es, das Chromosom 18 mit Fokus auf die Region des QTL mit Hilfe verschiedenster moderner Analyseverfahren erneut zu analysieren. Unter Einbeziehung der gewonnenen Ergebnisse erwies sich die Zielregion als weit komplexer als bisher angenommen. Im folgenden Kapitel wurden daher denkbare Strukturvarianten am Ziel QTL unter Einbeziehung bereits publizierter Erkenntnisse diskutiert.

5.1. Bestätigung des signifikanten Quantitativen Trait Locus

In einem vergangenen Forschungsprojekt der Arbeitsgruppe Populationsgenomik beschäftigten sich bereits MÜLLER et al. (2017) mit dieser Thematik. Aufgrund der Tatsache, dass sowohl in der Studie von MÜLLER et al. (2017) als auch in der vorliegenden Studie ein ähnliches Datenset und dieselbe Methode zur QTL Kartierung verwendet wurde, erfolgte zunächst ein Vergleich der Ergebnisse zwischen den beiden Studien. In der durchgeführten *cLDLA* von MÜLLER et al. (2017) wurde die signifikanteste Assoziation mit pKV und pTG ($LRT_{max} = 160{,}92$), bei 58.905.582 bp auf dem BTA18 lokalisiert. Das entsprechende Konfidenzintervall, welches mit einer Sicherheit von 95 % (KI95%) den

signifikanten QTL beinhaltete, reichte von 58.343.462 – 59.937.789 bp. Die vermutete Region des QTL von MÜLLER et al. (2017) konnte mit den Ergebnissen der vorliegenden Studie erneut bestätigt werden.

Das LRT Maximum für das Merkmal paternaler Kalbeverlauf wurde in der vorliegenden Studie bei 58.860.538 bp detektiert ($LRT_{max} = 122{,}9$). Das KI95% konnte auf einen Bereich von ca. 1,09 Mbp zwischen 58.343.346 und 59.432.662 bp festgesetzt werden. Nichtsdestotrotz befanden sich die Peakpositionen der beiden Studien 45.044 bp voneinander entfernt, was auf die minimalen Anpassungen des Studiendesigns zurückgeführt wurde. Das bestehende Tierset von MÜLLER et al. (2017) wurde für die vorliegende Studie aktualisiert und zusätzlich mit weiblichen Tieren ergänzt. Die höhere Anzahl an Haplotypen ermöglichte folglich eine exaktere Berechnung der LRT-Werte. MÜLLER et al. (2017) verwendeten außerdem in ihrer Varianzkomponentenanalyse ein gemischt lineares Model mit einer UAR Matrix aufgeteilt auf 110 Hauptkomponenten. Im Gegensatz dazu wurde in dieser Studie die vollständige UAR Matrix berücksichtigt, wodurch zufällige polygene Effekte besser in die Analyse miteinbezogen werden konnten. Darüber hinaus wurde das Markerset an das aktuelle *Bos taurus* Assembly ARS-UCD1.2 angepasst. Markerpositionen, die nicht in diesem Assembly lokalisiert werden konnten oder die Filterparameter nicht mehr erfüllten, wurden folglich vom Markerset ausgeschlossen. Die Liste der ausgeschlossenen Marker der Markersets von MÜLLER et al. (2017) und der vorliegenden Studie unterschieden sich zudem dadurch, dass manche Marker z.B. aufgrund der Inklusion weiterer Tiere in der vorliegenden Studie die MAF Grenze über- oder unterschritten. In Folge wurden betroffene Marker nicht für die weiteren Analysen verwendet. Des Weiteren konnte eine wesentliche Verbesserung der Haplotypisierung und Imputation durch die Verwendung des Programms *BEAGLE 5* anstelle *BEAGLE 3* und durch eine erheblich größere Anzahl an Paaren und Trios im verwendeten Basisdatensatz (d.h. alle verfügbaren Rinder, die mit dem gleichen SNP-Chip genotypisiert wurden) erzielt werden.

Auch in der vorliegenden Studie konnte erneut der erstmals von COLE et al. (2009) publizierte QTL bestätigt werden, jedoch lag der signifikanteste SNP rs109478645 (57.137.302 bp) 1,7 Mbp von der hier detektierten Peakposition entfernt. Eine ähnliche Abweichung stellten auch MÜLLER et al. (2017) zu

ihren Ergebnissen fest und führten dies auf die unterschiedlichen Methoden der QTL-Kartierung zurück. Während in den Studien der Arbeitsgruppe Populationsgenomik eine kombinierte Kopplungsungleichgewichts- und Kopplungsanalyse verwendet wurde, wählten COLE et al. (2009) eine genomweiten Assoziationsstudie. Die in der cLDLA rekonstruierten Haplotypen wiesen im Gegensatz zur GWAS eine geringere Sensitivität gegenüber fehlenden Markern auf und gewährleisteten folglich eine präzisere Bestimmung des QTL.

In dieser Studie wurde zur Detektion des QTL als phänotypisches Merkmal der relative Zuchtwert pKV herangezogen. EKINE et al. (2014) beschrieben in ihrer Publikation vermehrt falsch-positive Ergebnisse der GWAS aufgrund der Verwendung geschätzter Zuchtwerte engverwandter Individuen. Um derart falsch-positive Ergebnisse möglichst gering zu halten, wurde in dieser Studie keine GWAS, sondern eine Haplotyp-gestützte cLDLA-Kartierung angewendet. In der hier durchgeführten cLDLA wurde zudem ein Vektor der zufällig polygenen Effekte in der Varianzkomponentenanalyse miteinbezogen (siehe Kapitel 3.2.3.4). Diese Herangehensweise wurde u.a. von JIANG et al. (2014) in ihrer Studie zu Milchproduktionsmerkmalen in chinesischen Holsteins empfohlen, um möglichst akkurate Ergebnisse zu erhalten. Darüber hinaus repräsentierte das verwendete Datenset ein komplexes Pedigree töchtergeprüfter Bullen mit hunderten von Nachkommen. Im Gegensatz dazu untersuchten EKINE et al. (2014) lediglich rund 1.010 Individuen, wodurch die Sicherheit der GWAS aufgrund des kleinen Datensets vermindert ist (SAHANA et al., 2014). Des Weiteren konnte eine Sicherheit von durchschnittlich 72 % des relativen Zuchtwerts pKV sichergestellt werden.

5.2. Detektion des vermeintlich kausalen Haplotypen

Erstmals identifizierten MÜLLER et al. (2017) einen kausalen Haplotypen, der mit den Merkmalen paternaler Kalbeverlauf und paternale Totgeburt in Deutschen Holsteins assoziiert werden konnte. Dieser Haplotyp stand in einem populationsweiten Kopplungsungleichgewicht mit dem „schädlichen" QTL-Allel Q und reichte von 54.814.615 – 59.937.789 bp. Mit Hilfe des Vergleichs der Allele von zusätzlichen HD-Markern konnte zudem der

Haplotyp auf einen Bereich von 57.941.736 – 58.442.683 bp weiter eingegrenzt werden. In der aktuellen Studie gelang es, allein durch die Verwendung von 50K Markerdaten die Region des kausalen Haplotypen zwischen 57.922.208 – 60.057.741 bp zu detektieren. Da HD Marker-Haplotypen lediglich für 256 der 2.697 Tiere zur Verfügung standen und eine unsichere Imputation insbesondere in chromosomalen Regionen mit einer niedrigen Markerdichte angenommen wurde, wurde auf eine weitere Eingrenzung des Haplotypen auf Basis von imputierten HD Marker-Haplotypen verzichtet. Die Ungenauigkeiten der Imputation treten vor allem dann auf, wenn die geschätzten fehlenden Marker-Haplotypen bei falschen Genotypen über- und bei richtigen Genotypen unterbewertet werden (BAND et al., 2013).

Des Weiteren deckte sich der Bereich des kausalen Haplotypen mit der signifikant abweichenden Region (58–60 Mbp) mehrerer durchgeführter Analysen der vorliegenden Studie. Zum einen konnte in diesem chromosomalen Segment eine signifikant geringere Dichte (11.840 bp) an HD-Markern ($P < 1\times10^{-4}$) im Vergleich zum gesamten Chromosom (3.381 bp) festgestellt werden (Abbildung 6). Zum anderen wurden im selben chromosomalen Abschnitt zahlreiche segmentale Duplikationen (Abbildung 8) sowie signifikant kurze Contigs (Abbildung 7 und Tabelle 8) lokalisiert. Daher wäre eine Imputation selbst bei einem größeren Anteil HD-typisierter Tiere äußerst unsicher. Infolgedessen wurde von einer weiteren Eingrenzung des kausalen Haplotypen mittels HD-Marker abgesehen. Die niedrigere Markerdichte in diesem Bereich wurde folgendermaßen erklärt. Um einen Markerchip zu erstellen, werden die Sequenzen der ausgewählten Proben gegen ein Referenzgenom derselben Spezies gemappt und auf Alternativallele untersucht. Die detektierten Alternativallele sowie die Verteilung der Referenz- und Alternativallele durchlaufen anschließend einen spezifischen Filterprozess, der an verschiedene Kriterien angepasst wird. Dabei finden allgemein gültige genetische Regeln, die Minor Allele Frequency (MAF), der Abstand zwischen Markern und bereits bekannte SNPs Beachtung (GROENEN et al., 2011; TOSSER-KLOPP et al., 2014). Marker, die nun eine hohe Qualität und einen hohen Informationsgehalt aufweisen, werden für die Erstellung des Markerchips priorisiert. Dieses SNP-Ranking wird vor allem durch komplexe Mutationsgeschehen negativ beeinflusst (BANKEVICH & PEVZNER, 2020). Programme, wie das GATK sind derzeit nicht dazu in der

Lage, zuverlässig Allelvarianten in komplexen Regionen zu identifizieren. Das bedeutet, dass in diesen Abschnitten weder eine Aussage darüber getroffen werden kann, um welche Mutationsformen es sich handelt, noch über die exakte Ausdehnung dieser. Zusätzlich wird die Komplexität durch die Folgen zelleigener Reparaturvorgänge der DNA weiter erhöht (PU et al., 2018; BANKEVICH & PEVZNER, 2020). Die Vermutung, dass die signifikant geringere Markerdichte im Bereich des kausalen Haplotypen mit der Präsenz komplexer struktureller Genomvarianten zusammenhängt, bestätigte sich durch die Identifikation segmentaler Duplikationen. In der Region von 58–60 Mbp konnte eine deutlich höhere Anzahl an SD im Vergleich zum gesamten Chromosom beobachtet werden (Abbildung 8). Bereits mehrere Autoren beschrieben ein erhöhtes Vorkommen an Copy Number Variationen am Ende des BTA18 (LIU et al., 2009; SEROUSSI et al., 2010; BICKHART et al., 2012; BOUSSAHA et al., 2015; ZHANG et al., 2015; GAO et al., 2017; KEEL et al., 2017). Außerdem identifizierte LIU et al. (2009) mittels Fluoreszenz in-situ Hybridisierung ebenfalls eine segmentale Duplikation am distalen Chromosomenende.

In der Untersuchung ausgewählter Tiere mittels *de novo* Assembly Techniken erhärtete sich der Verdacht eines komplexen Mutationsgeschehens in diesem Chromosomensegment. Innerhalb der 2 Mbp von 58–60 Mbp konnte kein Contig eine Länge von mehr als 10 kb erreichen, obwohl für das gesamte Genom eine hohe mittlere Contig N50 von über 6 Mbp erreicht wurde (Abbildung 7 und Tabelle 8). Das Auftreten kurzer Contigs in diesem Abschnitt konnte zusätzlich durch das Mapping der Contigs an das UOA_Anugs_1 anstelle des ARS-UCD1.2 Referenzgenoms vergleichbar reproduziert werden (Abbildung 7). In Folge komplexer chromosomaler Genomvarianten können vermehrt Fehler in der Sequenzierung und im nachfolgenden Assembling auftreten (VOLLGER et al., 2019). Mit der Oxford Nanopore Long-Reads Sequenzierungstechnik ist es zwar möglich einfache strukturelle Variationen wie Translokationen, Tandem-Duplikationen und einfache Inversionen zu identifizieren. Handelt es sich allerdings um komplexe chromosomale Aberrationen, nimmt die Qualität des Nanoporesignals kontinuierlich ab (SPEALMAN et al., 2019). Algorithmen von Basecallern wie *GUPPY* (PAYNE et al., 2020) werden verwendet, um die Spannungsunterschiede beim Durchtritt der DNA durch die Nanopore in eine Basensequenz zu übersetzen. Trifft der Basecaller auf komplexe Strukturen, wird die Genauigkeit, mit welcher die

Basen in diesen DNA-Abschnitten korrekt ausgelesen werden, reduziert oder der Lesevorgang abgebrochen. Als Folge enthalten die generierten Long-Reads in den betroffenen Abschnitten gar keine Basen oder vermehrt falsche Basen bei gleichzeitig sinkendem Phred-Score (SPEALMAN et al., 2019). Wie in Kapitel 3.2.4 beschrieben, wird eine einzelsträngige DNA durch eine Nanopore von der negativen *cis*- zur positiven *trans*-Seite transportiert. Segmentale Duplikationen sowie ähnliche komplexe Chromosomenaberrationen begünstigen die Bildung von Sekundär- und Tertiärstrukturen sowohl auf der *cis*- als auch auf der *trans*-Seite. Diese Strukturen können den DNA-Transport sowie die Ableitung der Basensequenz erschweren oder sogar komplett verhindern. Bei der anschließenden Verwendung von Assembling-Programmen werden dann aufgrund der fehlenden oder kurzen bzw. falschen Sequenzen nur kurze Contigs erzeugt. Hinzu kommt, dass eingesetzte Assembler Paraloge Allele, d.h. identische Sequenzen aufgrund SD, zusammenlagern und daraufhin Lücken im Chromosom entstehen (VOLLGER et al., 2019).

Zusammenfassend zeichnete sich die relativ große Region des vermeintlich kausalen Haplotypen durch eine signifikant geringe Markerdichte, eine Vielzahl segmentaler Duplikationen sowie kurzer Contigs aus. In Folge der geringen Markerdichte sowohl auf dem 50K als auch dem HD-Chip erwies sich die Imputation bzw. Ableitung der Marker-Haplotypen in diesem Segment als besonders unsicher, wodurch keine Möglichkeit einer weiteren Eingrenzung des Haplotypen selbst durch Erweiterung des HD Tiersets bestand. Der Nachweis segmentaler Duplikationen und kurzer Contigs lieferte zudem Hinweise auf ein komplexes Mutationsgeschehen in diesem chromosomalen Abschnitt, welches erheblich die Ausführung von Sequenzierungs- und Assemblierungstechniken negativ beeinflusst. Dementsprechend sind Analysen wie die Kartierung von QTL und die nachfolgende Detektion möglicher kausaler Mutationen, in derart komplexen chromosomalen Segmenten nicht genauso präzise durchführbar, wie dies z.B. in Bereichen mit hoher Markerdichte der Fall wäre.

5.3. Die 16-Kb Lücke als kausales Mutationsgeschehen

Die visuelle Untersuchung von vier ausgewählten Long-Read HF Sequenzen in IGV ergab eine identische Region von 16.000 bp (58,846,130 – 58,861,709 bp), die durch das vollständige Fehlen gemappter Reads hervorstach. Darüber hinaus befand sich diese 16-Kb Lücke nahe der zentralen Position des KI95% und beinhaltete das *LRTmax* bei 58.860.538 bp (Anhang 2). Unter Berücksichtigung eines Genotypen-abhängigen Musters der DNA-Fragmente konnte die 16-Kb Lücke als kausale Mutation assoziiert mit dem Merkmal pKV definitiv ausgeschlossen werden. Die untersuchten heterozygoten HF Kühe (DHFnano01 und DHFnano02) trugen jeweils nur eine Kopie des kausalen Haplotypen. Das bedeutet, dass innerhalb der 16-Kb Lücke neben Reads die von der Deletion betroffen sind, erfolgreich mappende Reads zu erwarten wären. Des Weiteren würden die beiden homozygoten HF Bullen ein vollkommen unterschiedliches Bild der gemappten Reads aufweisen. Die Probe DHFnano03 (Gruppe Q/Q) wäre durch die vollständige Deletion aller Reads gekennzeichnet, da in diesem Bullen sowohl der maternale als auch der paternale Haplotyp das kausale Haplotypenmuster aufweisen. Im Gegensatz dazu wäre beim Bullen DHFnano04 keines der DNA-Fragmente von der Deletion betroffen, da der homozygote Haplotyp dieser Probe (Gruppe q/q) nichts mit dem kausalen Haplotypen gemeinsam hatte. Jedoch wiesen alle vier Proben unabhängig von den jeweiligen Haplotypen ein vergleichbares Bild fehlender Reads auf. Hinzu kommt, dass eine kausale Mutation ausschließlich in dieser Rasse auftreten würde, da der signifikante QTL bisher nur in reinrassigen Holstein und Veredelungskreuzungen mit diesen nachgewiesen wurde. Jedoch wiesen alle Proben unabhängig von der Rasse eine nahezu identische 16-Kb Lücke auf (Anhang 3 und 4).

Um die Ursache für das Fehlen der Reads aufzuklären, wurde die 16-Kb Lücke genauer untersucht. Durch den Nachweis von Reads mit *linked supplementary Alignments* innerhalb der 16-Kb Lücke konnten weitere Beweise einer weitreichenden komplizierten Genomvariante gesammelt werden. Hierbei werden Reads markiert, die durch Mappingprogramme getrennt und somit lückenhaft an das Referenzgenom gemappt wurden. Derartig abweichende Reads sind insbesondere in Regionen mit komplexen chromosomalen Aberrationen zu beobachten (PACIFIC BIOSCIENCES, 2017). Des Weiteren wiesen nur jeweils 11 % der Reads in den heterozygoten Proben DHFnano01 und DHFnano02 als auch in der Probe des Kärntner

Blondvieh *linked supplementary Alignments* auf. Im Falle einer kausalen Mutation der 16-Kb Lücke wäre auch hier eine Genotypen abhängige Verteilung der Reads mit und ohne *linked supplementary Alignments* zu erwarten. Demnach wären in 50% der Reads in den heterozygoten Proben, in allen Reads der homozygoten Probe DHFnano03 bzw. in keinem Read der Proben DHFnano04, KBVnano05 und FVnano06 *linked supplementary Alignments* nachweisbar. Da keines dieser Kriterien zutraf, konnte bestätigt werden, dass es sich bei der 16-Kb Lücke nicht um eine HF spezifische Mutation des Kalbekomplexes handelte.

Im Zuge des Validierungs-Alignments der heterozygoten HF Kühe an das UOA_Angus_1 Assembly konnten zunächst Long-Reads detektiert werden, welche die konvertierte Region der 16-Kb Lücke überdeckten. Jedoch befand sich innerhalb des KI95% des QTL (6.143.282 – 7.212.507 bp) zwischen 6.716.809 und 6.717.310 bp (d.h. 501 bp) eine ähnliche Lücke ohne jegliche Reads (Abbildung 5). Des Weiteren wurden 15 Marker, die innerhalb des konvertierten KI95% des QTL lagen, an einer weiteren Position im Angusgenom lokalisiert. Mehrere Positionen von ein und demselben Marker deuten darauf hin, dass die Marker dupliziert wurden. Folglich können diese SNPs an zwei oder mehr Positionen in einem Genom detektiert werden (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2020). Befinden sich wie in der Angus Referenz die mehrdeutigen SNPs in unmittelbarer Nachbarschaft zueinander, kann angenommen werden, dass auch die Region zwischen den Markern dupliziert wurde. Das bedeutete, dass jeweils die Blöcke mit gleichen Markern an zwei verschiedenen Positionen die Kriterien einer segmentalen Duplikation erfüllten (Basenkongruenz von mehr als 90 % und mindestens 5 kb lang (SHARP et al., 2005)). In Folge war eine eindeutige Zuordnung des KI95% im Angusgenom sowie die Identifikation von Kandidatengenen bzw. kausalen Mutationen nicht möglich. Des Weiteren würden die mehrdeutigen Marker auf dem UOA_Angus_1 Assembly eine erneuten cLDLA erheblich negativ beeinflussen, wodurch keine präzise Kartierung des QTL durchführbar wäre. Infolgedessen, dass die Marker auf dem ARS-UCD1.2 Assembly eindeutig an eine Position lokalisiert werden konnten, erwies sich dieses Assembly zur Durchführung der Analysen als besser geeignet.

Durch die Analyse der 16-Kb Lücke konnte die in Kapitel 5.2 aufgestellte Hypothese einer Anhäufung komplizierter Genomvarianten manifestiert werden. Des Weiteren bestätigte der fehlende Nachweis struktureller Varianten in ONT und Illumina sequenzierten Holsteinproben, dass vermutlich keine einfachen Strukturvarianten (z.B. einfache Inversionen, Deletionen oder Duplikationen) die genetischen Grundlagen des Kalbekomplexes beeinflussen. In den 21 Illumina sequenzierten Tieren konnte trotz Anwendung dreier verschiedener Programme (*LUMPY*, *BREAKDANCER* und *MANTA*) keine strukturelle Variante im Bereich des kausalen Haplotypen lokalisiert werden. In den vier ONT sequenzierten HF Proben detektierte das Programm *NANOVAR* zwar 33 SV innerhalb der Zielregion (Tabelle 10), diese erfüllten jedoch nicht die Voraussetzungen einer kausalen Mutation. Zweiunddreißig SV wurden gleichermaßen in allen Proben unabhängig ihres Haplotypen lokalisiert, wodurch eine Assoziation mit dem kausalen Haplotypen widerlegt werden konnte. Lediglich eine Deletion trat nur in den Trägern des kausalen Haplotypen auf, jedoch konnte auch diese Deletion als kausale Mutation ausgeschlossen werden, da sie auch in den heterozygoten Proben in homozygoter Form vorlag.

Darüber hinaus stellten mehrere Studien komplexe chromosomale Abweichungen am distalen Ende des Chromosoms fest. BICKHART et al. (2012) und BOUSSAHA et al. (2015) detektierten mehrere Insertionen, Deletionen und Tandem-Duplikationen in unmittelbarer Nähe des Quantitiativen Trait Locus. Des Weiteren stimmte eine Tandem-Duplikation von 58.820.108 – 59.381.142 bp (BOUSSAHA et al., 2015) und eine Insertion von 58.742.045 – 59.010.707 bp (BICKHART et al., 2012) exakt mit der Region der 16-Kb Lücke überein. IANNUZZI et al. (2020) untersuchten in einer weiteren Kartierungsstudie Störungen in der Geschlechtsentwicklung (Disorder of Sexual Developments (DSD)) von männlichen Wiederkäuern. Durch die kombinierte Anwendung von zytogenetischen Analysen und Ganzgenom Sequenzierungsmethoden konnten weitere Befunde der strukturellen Besonderheiten des bovinen Chromosom 18 erhoben werden. Die Autoren identifizierten eine Tandem-Fusion-Translokation ausgehend von den Telomeren des BTA18 zu dem Zentromer des Chromosom 27. Durch diese Translokation sind in Folge Teile der DNA am distalen Ende des BTA18 verloren gegangen. Die detektierten strukturellen Varianten von IANNUZZI et al. (2020) lagen zwar außerhalb des kausalen Haplotypen dieser Studie,

unterstützen jedoch die Hypothese eines komplexen Mutationsgeschehens am distalen Strangende des Chromosoms.

Jedoch ist zu beachten, dass neben genetischen Effekten auch epigenetische Faktoren Einflüsse auf die Gene des Kalbekomplexes nehmen können. Wie in Kapitel 2.6 beschrieben, führen Histonmodifikationen bzw. DNA-Methylierungen zu einer Aktivierung bzw. Inaktivierung von Genen (JAENISCH & BIRD, 2003; JANNING & KNUST, 2004). FANG et al. (2019) wiesen bei der Untersuchung spermaler DNA im zweiten Intron (BTA18: 57.777.046 – 57.807.070 bp) des *Zink Finger Protein 613* Gens eine differentiell methylierte Region (DMR) nach, die mit den Merkmalen Trächtigkeitsdauer, Schwergeburten, Körpergröße und Konzeptionsrate in Holstein-Friesian Tieren assoziiert werden konnte. Die Autoren stellten zudem fest, dass Tiere mit einer verlängerten Trächtigkeitsdauer, vermehrten Schwergeburten und größeren Kälbern eine höhere Methylierungsrate in speziell dieser differentiell methylierten Region aufwiesen. Mit dem Anhängen oder Entfernen einer Methylgruppe an CpG Dinukleotide vor einem Initationscodon werden die zugrundeliegenden Gene aktiviert oder inaktiviert (GRAW, 2006). Das bedeutet, durch DNA-Methylierungsprozesse im zweiten Intron könnte das *ZNF613* Gen inaktiviert werden. Nachfolgend wird durch das Fehlen des Transkriptionsfaktors ZNF613 die Genexpression aufgrund einer fehlerhaften Transkription verändert. Diese Gene könnten wiederrum Auswirkungen auf das Kalbeverhalten haben (siehe Kapitel 5.4). Nichtsdestotrotz sind weitere funktionelle Untersuchungen notwendig, um einen eindeutigen Zusammenhang zwischen dem *ZNF613* Gen und dem Kalbeverhalten zu bestätigen.

5.4. Potenzielle Kandidatengene

Insgesamt konnten vier SNPs in der Region des vermeintlich kausalen Haplotypen lokalisiert werden, deren Referenz- und Alternativallele mit der Verteilung des kausalen Haplotypen in den sequenzierten Tieren übereinstimmten (Tabelle 9). FANG et al. (2019) und PURFIELD et al. (2020) beschrieben in ihren Publikationen ebenfalls eine signifikante Assoziation des SNP rs381577268 (57.816.137 bp) mit verschiedenen Fruchtbarkeits- (hier v.a. Trächtigkeitsdauer) und Körperkonditionsmerkmalen. Darüber hinaus

konnte das nahegelegene *Zink Finger Protein 613* (*ZNF613*) Gen (57.774.874. – 57.816.078 bp) als wichtiges Kandidatengen identifiziert werden. Mutationen bzw. Veränderungen der DNA-Methylierung, die dieses Gen betreffen, könnten in Folge den Kalbeverlauf negativ beeinflussen. Der pathologische Mechanismus dahinter basiert auf dem fortschreitenden Wachstum des ungeborenen Kalbes aufgrund der verlängerten Trächtigkeitsdauer. Die in Relation zum Muttertier zu großen Kälber fördern nachfolgend das Auftreten von Störungen bei der Geburt (RICHTER & GÖTZE, 1993). Einen ähnlichen Mechanismus beschrieben sowohl COLE et al. (2009) als auch MAO et al. (2016) im Zusammenhang mit den beiden Genen *SIGLEC-13* (57.136.157 – 57.142.779 bp) und *CD33* (57.122.286 – 57.131.819 bp), welche wie das humane *SIGLEC-6* Gen der *SIGLEC*-Genfamilie angehören. Dem *SIGLEC-6* Gen konnte ein direkter Einfluss auf die Gestationslänge nachgewiesen werden. Die Geburt wird durch ein Zusammenspiel vieler verschiedener biochemischer Prozesse eingeleitet, darunter die Bindung von Leptinmolekülen durch Genprodukte des *SIGLEC-6* Gens. Sind nun die Genprodukte aufgrund von Mutationen nicht mehr dazu in der Lage, Leptin zu binden, ist einer der Einleitungsmechanismen gestört und die Geburt verzögert sich (BRINKMAN-VAN DER LINDEN et al., 2007). Die Autoren nahmen an, dass auf die gleiche Weise eine Mutation des *SIGLEC-13* bzw. *CD33* Gens eine mangelhafte Leptinbindung der Genprodukte auslöst und so die Trächtigkeitsdauer verlängert wird. Darüber hinaus konnte eine Korrelation zu Körperkonditionsmerkmalen hergestellt werden. Treten nun eine verlängerte Trächtigkeitsdauer in Kombination mit großen Kälbern auf, steigt die Gefahr einer Schwergeburt rapide an (COLE et al., 2009; MAO et al., 2016). Jedoch konnte in dieser Studie weder der SNP rs109478645 (57.137.302 bp) (COLE et al., 2009) noch der von MAO et al. (2016) detektierte SNP rs136283363 (57.089.460 bp) in den ONT sequenzierten Holsteinproben nachgewiesen werden. In der Nähe der beiden Top SNPs rs464221818 (59.329.179 bp) und rs472502785 (59.345.689 bp) konnte das *Zink Finger Protein 665-like* (*ZNF665-like*) Gen (59.375.765 – 59.378.104 bp) lokalisiert werden. ZHANG et al. (2016) stellten außerdem eine Assoziation dieser beiden SNPs zu dem Merkmal Lebensdauer in Holstein Kühen fest. In unmittelbarer Nähe des Kandidaten-SNPs rs381878735 (59.574.329 bp) befanden sich zwei Zink Finger Protein Gene, das *ZNF677-like* (57.440.901 – 59.536.515 bp) und das *ZNF665 Gen* (59.416.313 – 59.431.145 bp). Zink

Finger Proteine zählen neben den Helix-Turn-Helix-, Homöodomänen- und Leucin-Zipper-Proteinen zu den Transkriptionsfaktoren, welche der Regulation und Initiation der Expression eukaryotischer Gene dienen. Mit Hilfe der Transkriptionsfaktoren wird eine exakte Anlagerung der RNA-Polymerase an das Initiationscodon der DNA ermöglicht. Wird dieser Vorgang durch Mutationen der Transkriptionsfaktoren kodierenden Gene gestört, wird aufgrund der fehlerhaften Initiation eine Leserasterverschiebung oder der Verlust bzw. das Anfügen weiterer Aminosäuren an die gebildete RNA ausgelöst. In Folge können fehlerhafte oder funktionsunfähige Proteine exprimiert werden (GRAW, 2006). Nichtsdestotrotz konnten bisher in keiner Studie eindeutige Beweise gesammelt werden, die einen Einfluss der *Zink Finger Proteine 665*, *665-like* und *677-like* auf den Kalbeverlauf bestätigen.

5.5. Einfluss komplexer chromosomaler Aberrationen auf polygene Merkmale und Identifikation kausaler Mutationen

Unter Berücksichtigung der bisherigen Erkenntnisse ist davon auszugehen, dass unabhängig von der untersuchten Rinderrasse am distalen Ende des Chromosom 18 ein oder mehrere komplexe chromosomale Aberrationen vorliegen. Jedoch wurde der signifikante QTL bisher ausschließlich in Holstein Tieren und Veredelungskreuzungen mit diesen nachgewiesen. Dies führte zu der Vermutung, dass in der kritischen Region zusätzlich zu dem bereits komplexen Mutationscluster ein oder mehrere zusätzliche Mutationen in HF Tieren mit dem kausalen Haplotypen auftreten. Allerdings wird die Identifikation einer Holstein-spezifischen Mutation wesentlich erschwert, da komplexe Aberrationen zu einer hohen Fehlerrate der Basensequenz im Bereich des QTL bzw. im gesamten Referenzgenom führen (siehe Kapitel 5.1) (SPEALMAN et al., 2019; VOLLGER et al., 2019). Des Weiteren können komplexe Mutationsformen wie segmentale Duplikationen selbst kausal für den abweichenden Phänotyp in HF Tieren sein. Die Basensequenz von Genen des Kalbekomplexes könnten aufgrund der von SD induzierten nicht-allelischen homologen Rekombination (NAHR) modifiziert werden (KOSZUL & FISCHER, 2009). In Folge wäre eine veränderte Wirkung der Gene bzw. ihrer Genprodukte zu erwarten (STANKIEWICZ & LUPSKI, 2002). Bei Betrachtung der Zuchtwerte von Bullen, die den homozygoten kausalen

Haplotypen aufweisen, sind diese deutlich niedriger als jene Werte von Bullen mit heterozygoten bzw. anderen nicht-kausalen Haplotypen. Jedoch werden die Kälber der homozygoten Bullen nicht automatisch alle tot- oder schwergeboren. Dementsprechend steigt nur die Wahrscheinlichkeit einer Schwergeburt bei Nachkommen von Bullen mit kausalen Haplotypen. Dies lässt sich unter anderem dadurch erklären, dass eine segmentale Duplikation nicht in jeder Meiose eines bzw. verschiedener Individuen eine nicht-allelische homologe Rekombination auslöst. Bei Bullen, die den kausalen Haplotypen aufweisen, ist das Auftreten von NAHR jedoch wahrscheinlicher, als dies bei Bullen ohne den kausalen Haplotypen der Fall wäre. Nichtsdestotrotz konnte aufgrund der hohen Anzahl segmentaler Duplikationen im Bereich des kausalen Haplotypen nicht eindeutig nachgewiesen werden, welche und wie viele SD im direkten Zusammenhang mit dem Kalbeverlauf stehen.

5.6. Ausblick

Das Ziel kausale Mutationen der Kandidatengene und damit den signifikanten QTL assoziiert mit dem Merkmal paternaler Kalbeverlauf in Deutschen Holsteins zu entschlüsseln, konnte mit den Ergebnissen der Studie nicht erreicht werden.

Die gewonnenen Ergebnisse dieser Studie tragen jedoch wesentlich dazu bei, die zugrundeliegende Struktur am Ziel QTL weiter aufzuklären. Um die kritische Region jedoch vollständig zu entschlüsseln, sind weitere Kenntnisse über die Mechanismen komplexer chromosomaler Aberrationen notwendig. Mit diesem Wissen können sowohl akkurate als auch kontinuierliche Basensequenzen generiert werden, die dazu beitragen, kausale Gene und Mutationen zu identifizieren. Die progressive Entwicklung der Long-Read Sequenzierungstechnik ermöglicht es schon jetzt, einfache strukturelle Varianten zu detektieren. Jedoch sind die derzeitig entwickelten Algorithmen der Basecaller bisher nicht dazu in der Lage vollständig, korrekte Sequenzen in Regionen komplizierter Genomvarianten zu produzieren.

Zur weiteren Beurteilung der identifizierten Kandidatengene ist eine gezielte Genotypisierung der SNPs in einem größeren Tierset zu empfehlen. Des Weiteren sollte der Aspekt epigenetischer Einflüsse auf potenzielle Kandidatengene nicht außer Acht gelassen werden. Die Identifikation methylierter DNA-Regionen könnte mit einer CRISPR/Cas9 zielorientierten Sequenzierung in Tieren mit extremen Phänotypen durchgeführt werden.

Bis zur vollständigen Aufklärung des bovinen Chromosom 18 und damit einhergehend des signifikanten QTL in Holstein-Friesian Rindern bedarf es noch weiterer Studien. Die Entwicklung von Methoden zur Aufklärung komplizierter struktureller Abweichungen ist daher von höchstem Interesse nicht nur in der Nutztiergenomik, sondern auch in der biologischen und humangenetischen Forschung.

6. Zusammenfassung

Im Rahmen dieser Arbeit wurde eine erneute Untersuchung des *Bos taurus* Autosom 18 mit dem Ziel durchgeführt, den genetischen Hintergrund des Kalbekomplexes in Holstein-Friesian Rindern aufzuklären. In vergangen Studien (u.a. COLE et al. (2009); MÜLLER et al. (2017); PURFIELD et al. (2020)) gelang es einen signifikanten QTL, der mit verschiedenen Fruchtbarkeits- und Körperkonditionsmerkmalen assoziiert ist, zwischen 50 und 60 Mbp auf dem Chromosom 18 zu lokalisieren. Darüber hinaus konnte nach unserem Kenntnisstand dieser QTL ausschließlich in reinrassigen Holstein Rindern und Veredelungskreuzungen mit dieser Rasse detektiert werden. Nichtsdestotrotz konnten bisher weder eindeutige Kandidatengene noch kausale Mutationen identifiziert werden, die mit dem Kalbekomplex eindeutig assoziiert werden konnten. Mit der Wahl der in dieser Studie ausgewählten Analysemethoden sollte daher eine erneute Untersuchung der Zielregion vorgenommen werden. Zu diesen Analysen zählten unter anderem eine kombinierte Kopplungsungleichgewichts- und Kopplungsanalyse (cLDLA), Third-Generation Sequenzierung, visuelle Untersuchung ausgewählter genomischer Sequenzen mit Hilfe des *Integrative Genomics Viewer*, Anwendung von *de novo* Assembly Techniken, sowie Methoden zur Identifikation segmentaler Duplikationen und Strukturvarianten.

Mit Hilfe der cLDLA gelang es, den signifikanten QTL in das aktuelle Referenzgenom einzubinden und zu bestätigen. Das *LRT*-Maximum von 122,9 wurde an der Position 58.860.538 bp mit dem entsprechenden Konfidenzintervall des QTL von 58.343.346 – 59.432.662 bp lokalisiert. Des Weiteren konnte der vermutlich kausale Haplotyp, assoziiert mit dem Merkmal paternaler Kalbeverlauf, auf einen Bereich zwischen 57.922.208 – 60.057.741 bp eingegrenzt werden. Bei genauerer Untersuchung des kausalen Haplotypen stellte sich heraus, dass die zugrundeliegende Struktur weit komplexer ist als bisher angenommen. Anhand der gewonnenen Ergebnisse kann angenommen werden, dass der Region zwischen 58 und 60 Mbp komplexe chromosomale Aberrationen, wie z.B. segmentale Duplikationen zugrunde liegen. Diese haben nicht nur einen negativen Einfluss auf die Ergebnisse verschiedener Analysen, sondern nach unserer Annahme auch auf den Kalbekomplex in Holstein-Friesian Kühen. Die aufgestellte Hypothese konnte anhand der folgenden Ergebnisse weiter

konkretisiert werden. Einerseits gelang es mittels Selbstkartierung des Chromosom 18 segmentale Duplikationen (SD) zu identifizieren, die im Vergleich zum gesamten Chromosom vor allem am distalen Ende des Chromosoms und innerhalb der 2 Mbp zwischen 58 und 60 Mbp vermehrt auftraten. Des Weiteren konnte eine signifikant geringe Markerdichte in diesem Bereich festgestellt werden, die vermutlich auf der Präsenz komplexer Genomvarianten beruht. Darüber hinaus zeigte die Untersuchung von ONT Long-Read sequenzierten Tieren mittels *de novo* Assembly Techniken, dass zwischen 58 – 60 Mbp die generierten Contigs eine Länge von 10 kb nicht überschritten und folglich in dieser Region keine durchgehende Basensequenz generiert werden konnte. Diese Ergebnisse konnten ebenfalls mit den negativen Auswirkungen komplizierter Strukturvarianten auf das Base-Calling der Sequenzierung und *de novo* Assembler erklärt werden.

Ferner zeigte die visuelle Untersuchung Illumina und ONT sequenzierter Proben eine Region von 16.000 bp, die durch das vollständige Fehlen gemappter Reads hervorstach. Diese 16-Kb Lücke überschnitt sich exakt mit dem LRT-Maximum, trat jedoch unabhängig von den Haplotypen der verschiedenen HF Proben und der eingesetzten Rasse auf. Somit konnte diese Lücke als kausale Mutation des Kalbekomplexes eindeutig ausgeschlossen werden, allerdings lieferte dieser Abschnitt weitere Hinweise auf ein komplexes Mutationsgeschehen. In drei Proben wiesen mehrere flankierende DNA-Fragmente der 16-Kb Lücke *linked supplementary Alignements* auf. Diese deuten darauf hin, dass die DNA-Fragmente durch die angewendeten Programme fehlerhaft and das Referenzgenom gemappt und nachfolgend fälschlicherweise separiert wurden. Bei Validierung der Mappingergebnisse mit dem UOA_Angus_1 Assembly als Referenzgenom konnten nicht nur eine vergleichbare Lücke innerhalb des Ziel QTL detektiert, sondern auch Auffälligkeiten bei Überprüfung der Markerpositionen festgestellt werden. Die Positionen von 15 SNPs in der Region des kausalen Haplotypen konnten im Angusgenom an mehr als einer Position konvertiert werden. Die mehrdeutigen Marker befanden sich zudem verteilt auf drei Blöcke in unmittelbarer Nachbarschaft zueinander und erreichten dabei eine durchschnittliche Länge von 24 kb. Folglich erfüllten die Abschnitte dieser drei Markerblöcke die Kriterien segmentaler Duplikationen, d.h. eine Basenkongruenz von mindestens 90 % und einer Sequenzlänge von mehr als 5 kb.

Im Rahmen einer Untersuchung der ONT sequenzierten Proben auf Kandidatengene oder kausale Mutationen konnten vier SNPs detektiert werden, die mit dem Merkmal paternaler Kalbeverlauf assoziiert werden konnten. Anhand der SNP Positionen wurden vier potenzielle Kandidatengene identifiziert, darunter das vielversprechende *ZNF613* Gen (FANG et al., 2019; PURFIELD et al., 2020), sowie zusätzlich die drei weiteren Gene *ZNF665*, *ZNF665-like* und *ZNF677-like*.

Die komplexe Struktur segmentaler Duplikationen führt zu weitreichenden Problemen in Sequenzierungs-, Assembling- und Mappinganalysen. Daher treten vermehrt abweichende und unsichere Ergebnisse in den betroffenen chromosomalen Segmenten auf, welche die Identifikation kausaler Mutationen erheblich beeinträchtigen. In einer systematischen Suche nach strukturellen Mutationen konnte keine eindeutige Assoziation mit dem Merkmal paternaler Kalbeverlauf nachgewiesen werden. Jedoch lieferten diese Ergebnisse einen indirekten Hinweis auf die genetischen Hintergründe des Kalbeverlaufs. Segmentale Duplikationen fördern das Auftreten nicht-allelischer homologer Rekombinationen (NAHR) und führen in Folge zu fehlerhaftem Crossing-Over während der Meiose. Die betroffenen Abschnitte des Chromatids bzw. Chromosoms sind gekennzeichnet durch Deletionen und Duplikationen bis hin zur Lebensunfähigkeit des Individuums. Diese Studie lieferte erste Anhaltspunkte, dass die identifizierten Kandidatengene oder deren regulatorischen Elementen möglicherweise aufgrund der NAHR modifiziert wurden und in Folge das Auftreten von Schwergeburten fördern könnten. Nichtsdestotrotz ist noch weitere Forschung notwendig, um die aufgestellte Hypothese eindeutig zu bestätigen.

7. Summary

The aim of this study was to re-examine the *Bos taurus* autosome 18 to clarify the genetic background of the calving complex in Holstein-Friesian cattle. Several studies (i.a. COLE et al. (2009), MÜLLER et al. (2017), PURFIELD et al. (2020)) located a significant QTL associated with paternal calving ease and stillbirth between 50 and 60 Mbp on BTA18. To our knowledge, this QTL is only present in purebred Holstein cattle and breeds upgraded with HF. Despite the economic and ethical importance of this QTL, none of the existing publications have been able to pinpoint candidate genes or causal mutations. In order to encrypt the target region, we used several different methods, e.g. combined linkage disequilibrium and linkage Analysis (cLDLA), Third-Generation sequencing, visual inspection of selected sequences with the *INTEGRATIVE GENOMICS VIEWER*, *de novo* assembly techniques, as well as different methods to identify segmental duplications and structural variations.

With the cLDLA, we were able to confirm and incorporate the position of the significant QTL on BTA18 in the current reference genome. The *LRT* maximum of 122.9 was located at 58,860,538 bp with a corresponding confidence interval from 58,343,346 – 59,432,662 bp. Moreover, we were able to narrow down a putative harmful haplotype associated with paternal calving ease to a region between 57,922,208 – 60,057,741 bp. We assumed that the region of the causal haplotype harbors an extensive, complex structure of chromosomal aberrations (e.g. segmental duplications) with a negative impact on the results of different analyses as well as on the calving complex in Holstein-Friesian cattle. This hypothesis could be confirmed by the obtained knowledge of this study. The selfalignement of chromosome 18 identified several segmental duplications (SD), which were present in an increased number at the distal end of the chromosome as well as between 58 and 60 Mbp. Furthermore, within this 2 Mbp region, the marker density was significantly decreased compared to the remaining chromosome, which was traced back to the presence of complex genomic variants. Beyond that, the contigs, which were generated during the *de novo* assembly analyses, did not reach a length of more than 10 kb between these 2 Mbp and therefore did not allow the reconstruction of a continuous sequence. We assumed that these findings were based on the negative effects that complex structural variations have on tools for base-calling and *de novo* assembling.

Additionally, an outstanding region of 16,000 bp was observed between 58,846,130 – 58,861,709 bp, which was marked by the total absence of mapped reads. This 16-Kb gap overlapped exactly with the LRT-maximum but occurred independently of the haplotypes of different HF samples and the tested breed. Consequently, we excluded this 16-Kb gap as the casual mutation. Still, the gap encouraged the assumption of complex mutational events as a possible cause of the observed QTL. In three samples, several flanking DNA fragments of the 16-Kb gap showed *linked supplementary alignments*. These indicate that the affected DNA fragments were incorrectly mapped to the reference genome by the used programs, which resulted in mistakenly separated reads. Apart from that we also located a comparable gap in a mapping study using the UOA_Anugs_1 assembly as reference and additionally observed several ambiguous SNP positions in the Angus genome. Within the region of the putative harmful haplotype the positions of 15 SNP could be converted more than once in the Angus genome. The markers were located adjacent to each other and could be allocated to three blocks with an average length of 24 kb. Taking the criteria of segmental duplications into account (i.e. base similarity of more than 90 % and a minimum length of more than 5 kb), the region of these three blocks can be assumed to harbor one or more segmental duplications.

An examination of ONT sequenced samples for candidate genes or causal mutations detected four SNPs that could be associated with paternal calving ease. Based on the SNP positions, four potential candidate genes were identified, including the promising *ZNF613* gene (FANG et al., 2019; PURFIELD et al., 2020) and the three additional genes *ZNF665*, *ZNF665-like* and *ZNF677-like*.

The complex structure of segmental duplications entails extensive difficulties in performing sequencing, *de novo* assembling and mapping analyses. Therefore, aberrant, and uncertain results in affected chromosomal segments occur more frequently and significantly hinder the identification of casual genes and their mutations. In a systematic search for structural mutations, no clear association with the traits paternal calving ease could be demonstrated. However, these results provided an indirect indication on the genetic background of the calving complex in HF cattle. Furthermore, segmental duplications cause atypical crossing-over events like non-allelic homologous

recombinations (NAHR). Therefore, affected chromosomal segments showed deletions and duplications that might even lead to nonviable individuals. This study provides first evidence that the detected candidate genes or their regulatory elements may have been modified due to NAHR events and subsequently affect the calving performance in Holstein-Friesian cattle. Nevertheless, further research is still required to confirm this hypothesis.

8. Literaturverzeichnis

Achimastou A. The enzymes making the cut in NGS library preparation. Boston, USA: GE Healthcare Life Sciences 2019: https://www.gelifesciences.com/en/us/news-center/enzymes-in-ngs-library-prep-10001. 30. März 2020.

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol 2020; 21: 30.

Ashwell MS, Heyen DW, Sonstegard TS, Van Tassell CP, Da Y, VanRaden PM, Ron M, Weller JI, Lewin HA. Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle. J Dairy Sci 2004; 87: 468-75.

Babraham Institute. FastQC. Cambridge, UK: 2017: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. 04. Oktober 2019.

Bailey JA, Eichler EE. Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 2006; 7: 552-64.

Band G, Le QS, Jostins L, Pirinen M, Kivinen K, Jallow M, Sisay-Joof F, Bojang K, Pinder M, Sirugo G, Conway DJ, Nyirongo V, Kachala D, Molyneux M, Taylor T, Ndila C, Peshu N, Marsh K, Williams TN, Alcock D, Andrews R, Edkins S, Gray E, Hubbart C, Jeffreys A, Rowlands K, Schuldt K, Clark TG, Small KS, Teo YY, Kwiatkowski DP, Rockett KA, Barrett JC, Spencer CC. Imputation-based meta-analysis of severe malaria in three African populations. PLoS Genet 2013; 9: e1003509.

Bankevich A, Pevzner P. mosaicFlye: Resolving long mosaic repeats using long error-prone reads. BioRxiv 2020;

Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res 2011; 21: 381-95.

Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, Song J, Schnabel RD, Ventura M, Taylor JF, Garcia JF, Van Tassell CP, Sonstegard TS, Eichler EE, Liu GE. Copy number variation of individual cattle genomes using next-generation sequencing. Genome research 2012; 22: 778-90.

Boussaha M, Esquerre D, Barbieri J, Djari A, Pinton A, Letaief R, Salin G, Escudie F, Roulet A, Fritz S, Samson F, Grohs C, Bernard M, Klopp C, Boichard D, Rocha D. Genome-Wide Study of Structural Variants in Bovine Holstein, Montbeliarde and Normande Dairy Breeds. PLoS One 2015; 10: e0135931.

Brinkman-Van der Linden EC, Hurtado-Ziola N, Hayakawa T, Wiggleton L, Benirschke K, Varki A, Varki N. Human-specific expression of Siglec-6 in the placenta. Glycobiology 2007; 17: 922-31.

Broad Institute. Picard Toolkit. Broad Institute, GitHub repository 2019: http://broadinstitute.github.io/picard/.

Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 2009; 84: 210-23.

Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet 2018; 103: 338-48.

Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007; 81: 1084-97.

Bundesanstalt für Landwirtschaft und Ernährung. Landwirtschaftliche Produktionswert 2020: Nach erster Schätzung rund 56,3 Milliarden Euro. Bundesanstalt für Landwirtschaft und Ernährung - BLE 2020: https://www.ble.de/SharedDocs/Pressemitteilungen/DE/2020/201216_Landwirtschaftlicher_Produktionswert.html#:~:text=Quelle%3A%20BLE-,Landwirtschaftliche%20Produktionswert%202020%3A%20Nach%20erster%20Sch%C3%A4tzung%20rund%2056%2C3%20Milliarden,)%20um%203%2C8%20Prozent. 11. Jänner 2021.

Bundesverband Rind und Schwein. Rassebeschreibung. Bonn: 2021: https://www.rind-schwein.de/brs-rind/population-2.html. 25.08.2021.

Burton PR, Tobin MD, Hopper JL. Key concepts in genetic epidemiology. Lancet 2005; 366: 941-51.

Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? Mol Cell 2013; 49: 825-37.

Castro CJ, Ng TFF. U(50): A New Metric for Measuring Assembly Output Based on Non-Overlapping, Target-Specific Contigs. J Comput Biol 2017; 24: 1071-80.

Chandran S. Understanding and adapting the generic hard-filtering recommendations. Cambridge, USA: Broad Institute, Massachusetts Institute of Technology 2016: https://gatkforums.broadinstitute.org/gatk/discussion/6925/understanding-and-adapting-the-generic-hard-filtering-recommendations. 25. Februar 2020.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature methods 2009; 6: 677-81.

Chen N, Fu W, Zhao J, Shen J, Chen Q, Zheng Z, Chen H, Sonstegard TS, Lei C, Jiang Y. BGVD: An Integrated Database for Bovine Sequencing Variations and Selective Signatures. Genomics Proteomics Bioinformatics 2020; 18: 186-93.

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 2016; 32: 1220-2.

Chou HH, Holmes MH. DNA sequence quality trimming and vector removal. Bioinformatics 2001; 17: 1093-104.

Cole JB, VanRaden PM, O'Connell JR, Van Tassell CP, Sonstegard TS, Schnabel RD, Taylor JF, Wiggans GR. Distribution and location of genetic effects for dairy traits. J Dairy Sci 2009; 92: 2931-46.

Cole JB, Wiggans GR, Ma L, Sonstegard TS, Lawlor TJ, Jr., Crooker BA, Van Tassell CP, Yang J, Wang S, Matukumalli LK, Da Y. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC Genomics 2011; 12: 408.

Collins A. Allelic association: linkage disequilibrium structure and gene mapping. Mol Biotechnol 2009; 41: 83-9.

Cunningham F, Achuthan P, Akanni W, Allen J, Amode M R, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, Cummins C, Davidson C, Dodiya KJ, Gall A, Girón CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Laird MR, Lavidas I, Liu Z, Loveland Jane E, Marugán JC, Maurel T, McMahon AC, Moore B, Morales J, Mudge JM, Nuhn M, Ogeh D, Parker A, Parton A, Patricio M, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sparrow H, Stapleton E, Szuba M, Taylor K, Threadgold G, Thormann A, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Yates AD, Zerbino DR, Flicek P. Ensembl 2019. Nucleic Acids Res 2018; 47: D745-D51.

Dallery JF, Lapalu N, Zampounis A, Pigné S, Luyten I, Amselem J, Wittenberg AHJ, Zhou S, de Queiroz MV, Robin GP, Auger A, Hainaut M, Henrissat B, Kim KT, Lee YH, Lespinet O, Schwartz DC, Thon MR, O'Connell RJ. Gapless genome assembly of Colletotrichum higginsianum reveals chromosome structure and association of transposable elements with secondary metabolite gene clusters. BMC Genomics 2017; 18: 667.

Dammann RH, Richter AM, Jiménez AP, Woods M, Küster M, Witharana C. Impact of Natural Compounds on DNA Methylation Levels of the Tumor Suppressor Gene RASSF1A in Cancer. Int J Mol Sci 2017; 18

Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. Nat Biotechnol 2016; 34: 518-24.

Deutscher Holstein Verband. German Holstein Genetics_Quality Progress Reliability. Bonn: Deutscher Holstein Verband 2019: https://www.rind-schwein.de/services/files/dhv/pdf/DHV_Brosch%C3%BCre_ZA_EN.pdf. 18.07.2019.

DeWan A, Klein RJ, Hoh J. Linkage disequilibrium mapping for complex disease genes. Methods Mol Biol 2007; 376: 85-107.

Ekine CC, Rowe SJ, Bishop SC, de Koning DJ. Why breeding values estimated using familial data should not be used for genome-wide association studies. G3 (Bethesda) 2014; 4: 341-7.

Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics 2015; 14: 305-14.

Ewing B, Green P. Base-calling of automated sequencer traces using phred. II.

Error probabilities. Genome research 1998; 8: 186-94.

Fang L, Jiang J, Li B, Zhou Y, Freebern E, Vanraden PM, Cole JB, Liu GE, Ma L. Genetic and epigenetic architecture of paternal origin contribute to gestation length in cattle. Commun Biol 2019; 2: 100.

Farnir F, Grisart B, Coppieters W, Riquet J, Berzi P, Cambisano N, Karim L, Mni M, Moisio S, Simon P, Wagenaar D, Vilkki J, Georges M. Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics 2002; 161: 275-87.

Feng X, Jiang J, Padhi A, Ning C, Fu J, Wang A, Mrode R, Liu JF. Characterization of genome-wide segmental duplications reveals a common genomic feature of association with immunity among domestic animals. BMC Genomics 2017; 18: 293.

Fürst C (2021) Zuchtwertschätzung beim Rind - Grundlagen, Methoden und Interpretationen. ZuchtData EDV-Dienstleistungen GmbH, Wien

Gao Y, Jiang J, Yang S, Hou Y, Liu GE, Zhang S, Zhang Q, Sun D. CNV discovery for milk composition traits in dairy cattle using whole genome resequencing. BMC Genomics 2017; 18: 265.

GATK-Team. ClipReads. Cambridge, USA Broad Institute, Massachusetts Institute of Technology 2019: https://gatk.broadinstitute.org/hc/en-us/articles/360036882971-ClipReads. 29. September 2020.

Gehrke LJ, Upadhyay M, Heidrich K, Kunz E, Klaus-Halla D, Weber F, Zerbe H, Seichter D, Graf A, Krebs S, Blum H, Capitan A, Thaller G, Medugorac I. A de novo frameshift mutation in ZEB2 causes polledness, abnormal skull shape, small body stature and subfertility in Fleckvieh cattle. Sci Rep 2020; 10: 17032.

Geldermann H (2005) Tier-Biotechnologie. 1. Aufl. Verlag Eugen Ulmer, Stuttgart

Georges M. Mapping, fine mapping, and molecular dissection of quantitative trait Loci in domestic animals. Annu Rev Genomics Hum Genet 2007; 8: 131-62.

Gilmour A, Gogel B, Cullis B, Thompson R (2009) Asreml User Guide Release 3.0. VSN International Ltd. Hemel Hemptead, HP1 1ES, UK

Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 2009; 10: 381-91.

Graw J (2006) Genetik. 4. Aufl. Springer-Verlag, Berlin Heidelberg

Groenen MA, Megens HJ, Zare Y, Warren WC, Hillier LW, Crooijmans RP, Vereijken A, Okimoto R, Muir WM, Cheng HH. The development and characterization of a 60K SNP chip for chicken. BMC Genomics 2011; 12: 274.

Grunert E, Bechtold M (1999) Fertilitätsstörungen beim weiblichen Rind. 3. Aufl. Enke, Stuttgart

Haley CS, Knott SA. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity (Edinb) 1992; 69: 315-24.

Hansen M, Lund MS, Pedersen J, Christensen L. Gestation length in Danish Holsteins has weak genetic associations with stillbirth, calving difficulty, and calf size. Livestock Production Science 2004; 91: 23-33.

Heuven HC, Bovenhuis H, Janss LL, van Arendonk JA. Efficiency of population structures for mapping of Mendelian and imprinted quantitative trait loci in outbred pigs using variance component methods. Genet Sel Evol 2005; 37: 635-55.

Holmberg M, Andersson-Eklund L. Quantitative trait loci affecting fertility and calving traits in Swedish dairy cattle. J Dairy Sci 2006; 89: 3664-71.

Huxley JN, Whay HR. Current attitudes of cattle practitioners to pain and the use of analgesics in cattle. Vet Rec 2006; 159: 662-8.

Iannuzzi A, Braun M, Genualdo V, Perucatti A, Reinartz S, Proios I, Heppelmann M, Rehage J, Hulskotter K, Beineke A, Metzger J, Distl O. Clinical, cytogenetic and molecular genetic characterization of a tandem fusion translocation in a male Holstein cattle with congenital hypospadias and a ventricular septal defect. PLoS One 2020; 15: e0227117.

Illumina Inc. Quality Scores for Next-Generation Sequencing - Assessing sequencing accuracy using Phred quality scoring. Illumina Inc. 2011: https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf. 18. Jänner 2021.

Illumina Inc. An Introduction to Next-Generation Sequencing Technology. 2015: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf. 27. März 2020.

Illumina Inc. BovineSNP50 v3 BeadChip. 2020a: https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_bovine_snp50.pdf. 14. Jänner 2021.

Illumina Inc. Differences Between NGS and Sanger Sequencing. 2020b: https://emea.illumina.com/science/technology/next-generation-sequencing/ngs-vs-sanger-sequencing.html. 27. März 2020.

Ilumina Inc. Data Sheet: Agrigenomics, BovineHD Genotyping BeadChip. 2015: https://support.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_bovineHD.pdf. 02.10.2019.

Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 2003; 33 Suppl: 245-54.

Janning W, Knust E (2004) Genetik. 1. Aufl, Georg Thieme Verlag, Stuttgart.

Jiang L, Liu X, Yang J, Wang H, Jiang J, Liu L, He S, Ding X, Liu J, Zhang Q. Targeted resequencing of GWAS loci reveals novel genetic variants for

milk production traits. BMC Genomics 2014; 15: 1105.

Joshi N, Fass J (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files

(Version 1.33). https://github.com/najoshi/sickle

Kaplan NL, Hill WG, Weir BS. Likelihood methods for locating disease genes in nonequilibrium populations. Am J Hum Genet 1995; 56: 18-32.

Keel BN, Keele JW, Snelling WM. Genome-wide copy number variation in the bovine genome detected using low coverage sequence of popular beef breeds. Anim Genet 2017; 48: 141-50.

Khaja R, MacDonald JR, Zhang J, Scherer SW. Methods for identifying and mapping recent segmental and gene duplications in eukaryotic genomes. Methods Mol Biol 2006; 338: 9-20.

Kolbehdari D, Wang Z, Grant JR, Murdoch B, Prasad A, Xiu Z, Marques E, Stothard P, Moore SS. A whole-genome scan to map quantitative trait loci for conformation and functional traits in Canadian Holstein bulls. J Dairy Sci 2008; 91: 2844-56.

Koszul R, Fischer G. A prominent role for segmental duplications in modeling eukaryotic genomes. C R Biol 2009; 332: 254-66.

Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 1999; 22: 139-44.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome research 2009; 19: 1639-45.

Kühn C, Bennewitz J, Reinsch N, Xu N, Thomsen H, Looft C, Brockmann GA, Schwerin M, Weimann C, Hiendleder S, Erhardt G, Medjugorac I, Forster M, Brenig B, Reinhardt F, Reents R, Russ I, Averdunk G, Blumel J, Kalm E. Quantitative trait loci mapping of functional traits in the German Holstein cattle population. J Dairy Sci 2003; 86: 360-8.

Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 2014; 15: R84.

Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. Bioinformatics 2002; 18: 452-64.

Lee C. Generating consensus sequences from partial order multiple sequence alignment graphs. Bioinformatics 2003; 19: 999-1008.

Lee SH, Van der Werf JH. Using dominance relationship coefficients based on linkage disequilibrium and linkage with a general complex pedigree to increase mapping resolution. Genetics 2006; 174: 1009-16.

Leroi AM, Bartke A, De Benedictis G, Franceschi C, Gartner A, Gonos ES, Fedei ME, Kivisild T, Lee S, Kartaf-Ozer N, Schumacher M, Sikora E, Slagboom E, Tatar M, Yashin AI, Vijg J, Zwaan B. What evidence is there for the existence of individual genes with antagonistic pleiotropic effects? Mech Ageing Dev 2005; 126: 421-9.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25: 2078-9.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009; 25: 1754-60.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018; 34: 3094-100.

Liu GE, Ventura M, Cellamare A, Chen L, Cheng Z, Zhu B, Li C, Song J, Eichler EE. Analysis of recent segmental duplications in the bovine genome. BMC Genomics 2009; 10: 571.

Low WY, Tearle R, Liu R, Koren S, Rhie A, Bickhart DM, Rosen BD, Kronenberg ZN, Kingan SB, Tseng E, Thibaud-Nissen F, Martin FJ, Billis K, Ghurye J, Hastie AR, Lee J, Pang AWC, Heaton MP, Phillippy AM, Hiendleder S, Smith TPL, Williams JL. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. Nat Commun 2020; 11: 2071.

Ma L, Cole JB, Da Y, VanRaden PM. Symposium review: Genetics, genome-wide association study, and genetic improvement of dairy fertility traits. J Dairy Sci 2019; 102: 3735-43.

Mao X, Kadri NK, Thomasen JR, De Koning DJ, Sahana G, Guldbrandtsen B. Fine mapping of a calving QTL on Bos taurus autosome 18 in Holstein cattle. J Anim Breed Genet 2016; 133: 207-18.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 2010; 20: 1297-303.

Medugorac I, Seichter D, Graf A, Russ I, Blum H, Gopel KH, Rothammer S, Forster M, Krebs S. Bovine polledness--an autosomal dominant trait with allelic heterogeneity. PLoS One 2012; 7: e39477.

Meuwissen TH, Goddard ME. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 2000; 155: 421-30.

Meuwissen TH, Goddard ME. Prediction of identity by descent probabilities from marker-haplotypes. Genet Sel Evol 2001; 33: 605-34.

Meuwissen TH, Karlsen A, Lien S, Olsaker I, Goddard ME. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 2002; 161: 373-9.

Meyers RA (2004) Encyclopedia of molecular cell biology and molecular medicine, 2 edn. Wiley-Backwell, Weinheim

Müller MP, Rothammer S, Seichter D, Russ I, Hinrichs D, Tetens J, Thaller G, Medugorac I. Genome-wide mapping of 10 calving and fertility traits in Holstein dairy cattle with special regard to chromosome 18. J Dairy Sci 2017; 100: 1987-2006.

National Center for Biotechnology Information. NCBI Genome Remapping

Service. Bethesda, Maryland, USA: National Center for Biotechnology Information, U.S. National Library of Medicine 2018: https://www.ncbi.nlm.nih.gov/genome/tools/remap. 05. September 2019.

National Center for Biotechnology Information. About our alignments - Assembly-Assembly Alignments. Bethesda, Maryland, USA: National Center for Biotechnology Information, U.S. National Library of Medicine 2020: https://www.ncbi.nlm.nih.gov/genome/tools/remap/docs/alignments. 08. September 2020.

National Center for Biotechnology Information (NCBI). Sequence Read Archive. Bethesda, Maryland, USA: National Center for Biotechnology Information, U.S. National Library of Medicine 2018a: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement. 25. August 2019.

National Center for Biotechnology Information (NCBI). SRA Toolkit Documentation. Bethesda, Maryland, USA: National Center for Biotechnology Information, U.S. National Library of Medicine 2018b: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc. 25. August 2019.

National Human Genome Research Institute. Talking Glossary of Genetic Terms - Contig. National Human Genome Research Institute 2011: https://www.genome.gov/genetics-glossary/Contig. 15. Jänner 2021.

Ng PC, Kirkness EF. Whole genome sequencing. Methods Mol Biol 2010; 628: 215-26.

Nicholas FW (2003) Veterinary Genetics. 2. Aufl. Blackwell Publishing Ltd, Carlton South Victoria

Oki M, Aihara H, Ito T. Role of histone phosphorylation in chromatin dynamics and its implications in diseases. Subcell Biochem 2007; 41: 319-36.

Olsen HG, Lien S, Svendsen M, Nilsen H, Roseth A, Aasland Opsal M, Meuwissen TH. Fine mapping of milk production QTL on BTA6 by combined linkage and linkage disequilibrium analysis. J Dairy Sci 2004; 87: 690-8.

Oxford Nanopore Technologies. How does nanopore DNA/RNA sequencing work? Oxford, UK: Oxford Nanopore Technologies 2020a: https://nanoporetech.com/how-it-works. 30. März 2020.

Oxford Nanopore Technologies. Media Gallery. Oxford, UK: Oxford Nanopore Technologies, 2020b: https://nanoporetech.com/about-us/for-the-media#image194757219&. 11. August 2020.

Oxford Nanopore Technologies. Types of nanopores. Oxford, UK: Oxford Nanopore Technologies, 2020c: https://nanoporetech.com/how-it-works/types-of-nanopores. 30. März 2020.

Paaby AB, Rockman MV. The many faces of pleiotropy. Trends Genet 2013; 29: 66-73.

Pacific Biosciences. IGV 3 Improves Support for PacBio Long Reads. Kalifornien, USA: Pacific Biosciences of California 2017: https://www.pacb.com/blog/igv-3-improves-support-pacbio-long-reads/.

15. Mai 2020.

Parker Gaddis KL, Null DJ, Cole JB. Explorations in genome-wide association studies and network analyses with dairy cattle fertility traits. J Dairy Sci 2016; 99: 6420-35.

Payne A, Holmes N, Rakyan V, Loose M. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. BioRxiv 2018: 312256.

Payne A, Holmes N, Clarke T, Munro R, Debebe B, Loose M. Nanopore adaptive sequencing for mixed samples, whole exome capture and targeted panels. bioRxiv. 2020;

Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. Long reads: their purpose and place. Hum Mol Genet 2018; 27: R234-r41.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D. Scaling accurate genetic variant discovery to tens of thousands of samples. BioRxiv 2018: 201178.

Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in complex trait studies. Nat Rev Genet 2010; 11: 800-5.

Precht M, Meier N, Tremel D (2004) EDV-Grundwissen: Eine Einführung in Theorie und Praxis der modernen EDV. 7. aktualisierte Aufl., Addison-Wesley Verlag, München

Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. Am J Hum Genet 2001; 69: 1-14.

Pu L, Lin Y, Pevzner PA. Detection and analysis of ancient segmental duplications in mammalian genomes. Genome research 2018; 28: 901-9.

Purfield DC, Bradley DG, Evans RD, Kearney FJ, Berry DP. Genome-wide association study for calving performance using high-density genotypes in dairy and beef cattle. Genet Sel Evol 2015; 47: 47.

Purfield DC, Evans RD, Berry DP. Breed- and trait-specific associations define the genetic architecture of calving performance traits in cattle. J Anim Sci 2020; 98.

R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing 2019: https://www.r-project.org/ 25. September 2019.

Richter J, Götze R (1993) Tiergeburtshilfe. 4. Aufl. Verlag Paul Parey, Berlin/Hamburg.

Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996; 273: 1516-7.

Rodríguez-Bermúdez R, Miranda M, Baudracco J, Fouz R, Pereira V, López-Alonso M. Breeding for organic dairy farming: what types of cows are needed? J Dairy Res 2019; 86: 3-12.

Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C, Hall R, Li W, Rhie A, Ghurye J, McKay SD, Thibaud-Nissen F, Hoffman J, Murdoch BM, Snelling WM, McDaneld TG, Hammond JA, Schwartz JC, Nandolo W, Hagen DE, Dreischer C, Schultheiss SJ, Schroeder SG, Phillippy AM, Cole JB, Van Tassell CP, Liu G, Smith TPL, Medrano JF. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience 2020; 9.

Rothwell NV (1993) Understanding Genetics : a mollecular approach, 1 edn. Wiley-Liss, Inc., New York.

Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nature methods 2019;

Sahana G, Guldbrandtsen B, Lund MS. Genome-wide association study for calving traits in Danish and Swedish Holstein cattle. J Dairy Sci 2011; 94: 479-86.

Sahana G, Guldbrandtsen B, Thomsen B, Holm LE, Panitz F, Brøndum RF, Bendixen C, Lund MS. Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. J Dairy Sci 2014; 97: 7258-75.

Schnabel RD, Sonstegard TS, Taylor JF, Ashwell MS. Whole-genome scan to detect QTL for milk production, conformation, fertility and functional traits in two US Holstein families. Anim Genet 2005; 36: 408-16.

Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes 2016; 9: 88.

Seidenspinner T, Bennewitz J, Reinhardt F, Thaller G. Need for sharp phenotypes in QTL detection for calving traits in dairy cattle. J Anim Breed Genet 2009; 126: 455-62.

Seroussi E, Glick G, Shirak A, Yakobson E, Weller JI, Ezra E, Zeron Y. Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs. BMC Genomics 2010; 11: 673.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 2005; 77: 78-88.

Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet 2013; 14: 204-20.

Spealman P, Burrell J, Gresham D. Nanopore sequencing undergoes catastrophic sequence failure at inverted duplicated DNA sequences. BioRxiv 2019;

Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders. Trends Genet 2002; 18: 74-82.

Terwilliger JD, Zollner S, Laan M, Paabo S. Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. Hum Hered 1998; 48: 138-54.

Terwilliger JD, Goring HH. Gene mapping in the 20th and 21st centuries:

statistical methods, data analysis, and experimental design. Hum Biol 2000; 72: 63-132.

Thaller G (2011) FUGATO-plus Projekt GENOTRACK. Christian Albrechts Universität zu Kiel, Kiel, Germany

Tham CY, Tirado-Magallanes R, Goh Y, Fullwood MJ, Koh BTH, Wang W, Ng CH, Chng WJ, Thiery A, Tenen DG, Benoukraf T. NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. Genome Biol 2020; 21: 56.

Thomasen JR, Guldbrandtsen B, Sorensen P, Thomsen B, Lund MS. Quantitative trait loci affecting calving traits in Danish Holstein cattle. J Dairy Sci 2008; 91: 2098-105.

Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013; 14: 178-92.

Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, Donnadieu-Tonon C, Eggen A, Heuven HC, Jamli S, Jiken AJ, Klopp C, Lawley CT, McEwan J, Martin P, Moreno CR, Mulsant P, Nabihoudine I, Pailhoux E, Palhiere I, Rupp R, Sarry J, Sayre BL, Tircazes A, Jun W, Wang W, Zhang W. Design and characterization of a 52K SNP chip for goats. PLoS One 2014; 9: e86227.

Van der Auwera GA. What s a VCF and how schould I interpret it ? Cambridge, USA: Broad Institute, Massachusetts Institute of Technology 2017: https://gatkforums.broadinstitute.org/gatk/discussion/1268/what-is-a-vcf-and-how-should-i-interpret-it. 27. Februar 2020.

van Ooijen JW. Accuracy of mapping quantitative trait loci in autogamous species. Theor Appl Genet 1992; 84: 803-11.

Vaser R, Sovic I, Nagarajan N, Sikic M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome research 2017; 27: 737-46.

Verband der Milcherzeuger Bayern e.V. Milchpreise in Deutschland und Bundesländer. München: Verband der Milcherzeuger Bayern e.V. (VMB) 2019: https://www.milcherzeugerverband-bayern.de/themen/rubrik-fuer-milcherzeuger/milchpreis/milchpreis-charts/. 06. November 2020.

Vereinigte Informationssysteme Tierhaltung w.V. Trend - Fakten - Zahlen 2019. Verden: Vereinigte Informationssysteme Tierhaltung w.V. 2019: https://www.vit.de/fileadmin/Wir-sind-vit/Jahresberichte/vit-JB2019-gesamt.pdf. 11. Mai 2020.

Vereinigte Informationssysteme Tierhaltung w.V. (2020) Ökonomische Kennzahlen für die Merkmale Kalbeverlauf und Totgeburten. Vereinigte Informationssysteme Tierhaltung w.V., Verden

Vignal A, Milan D, SanCristobal M, Eggen A. A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol 2002; 34: 275-305.

Visscher PM, Goddard ME. Prediction of the confidence interval of quantitative

trait Loci location. Behav Genet 2004; 34: 477-82.

VIT (2020) Trends - Fakten - Zahlen 2020. Vereinigte Informationssysteme Tierhaltung w.V., Verden

VIT (2021) Beschreibung der Zuchtwertschätzung für alle Schätzmerkmale bei den Milchrinderrassen für die vit mit der Zuchtwertschätzung beauftragt ist. Vereinigte Informationssysteme Tierhaltung w.V., Verden

Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, Graves-Lindsay TA, Wilson RK, Chaisson MJP, Eichler EE. Long-read sequence and assembly of segmental duplications. Nature methods 2019; 16: 88-94.

Von Engelhardt W, Breves G (2009) Physiologie der Haustiere. 3. Aufl., Enke Verlag, Stuttgart.

Weckselblatt B, Rudd MK. Human Structural Variation: Mechanisms of Chromosome Rearrangements. Trends Genet 2015; 31: 587-99.

Weller JI (2001) Quantitative Trait Loci Analysis in Animals. 1. Aufl. CABI Publishing Wallingford, United Kingdom

WHFF (2018) 2018 Annual Statistics Report - World. World Holstein Friesian Federation

Wick R. Porechop. Available online at: https://github.com/rrwick/Porechop: 2018: 25. April 2020.

Williams GC. Pleiotropy, natural selection, and the evolution of senescence. evolution 1957; 11: 398-411.

Womack JE (2012) Bovine genomics. 1st edition, Wiley-Blackwell, Oxford, United Kingdom.

Wu R, Zeng ZB. Joint linkage and linkage disequilibrium mapping in natural populations. Genetics 2001; 157: 899-909.

Zhang Q, Ma Y, Wang X, Zhang Y, Zhao X. Identification of copy number variations in Qinchuan cattle using BovineHD Genotyping Beadchip array. Mol Genet Genomics 2015; 290: 319-27.

Zhang Q, Guldbrandtsen B, Thomasen JR, Lund MS, Sahana G. Genome-wide association study for longevity with whole-genome sequencing in 3 cattle breeds. J Dairy Sci 2016; 99: 7289-98.

Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol 2000; 7: 203-14.

Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, Marcais G, Roberts M, Subramanian P, Yorke JA, Salzberg SL. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol 2009; 10: R42.

9. Anhang

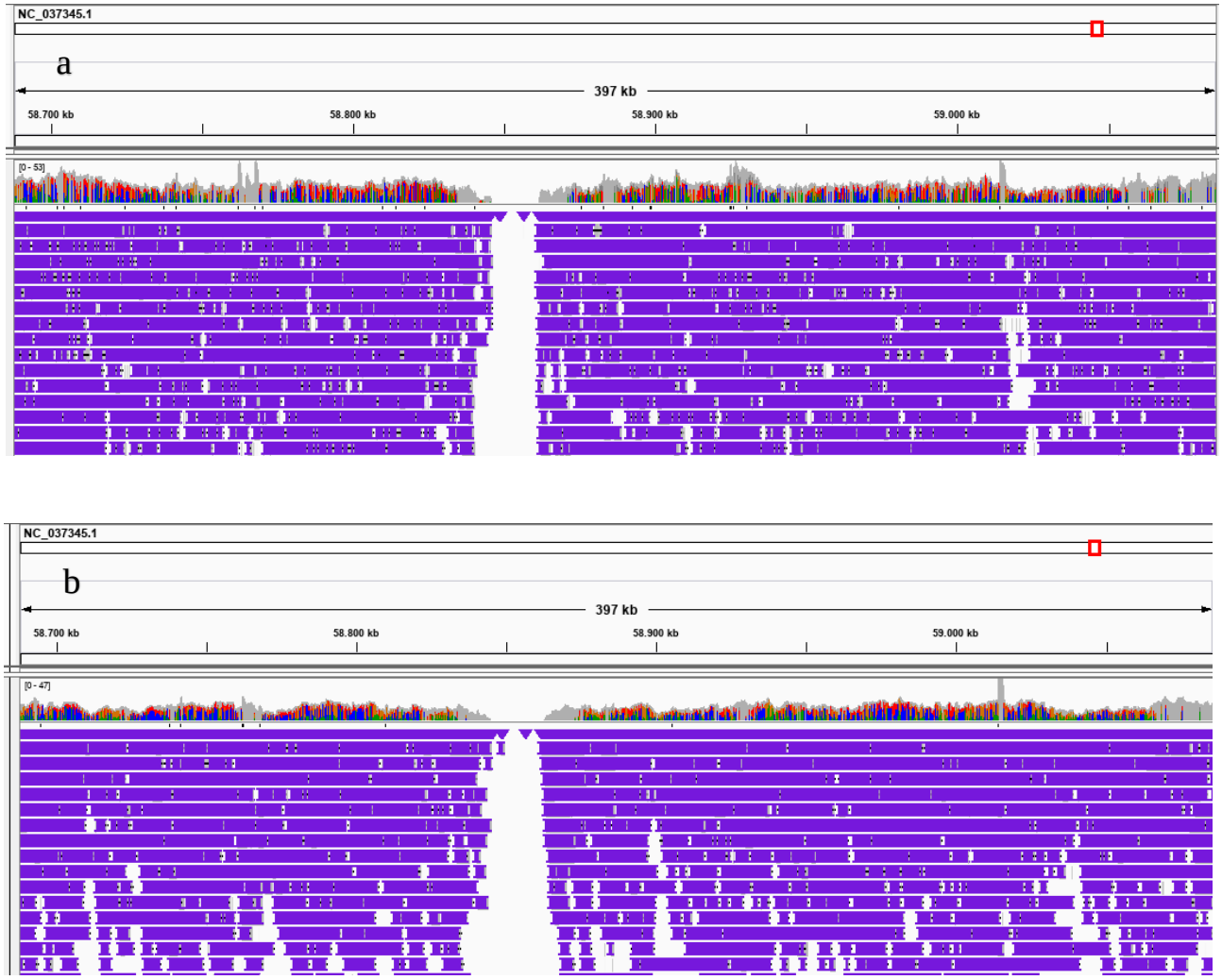
Anhang 1: Qualitätsstatistik der zusätzlichen 21 paired Illumina Short-Read Holstein Sequenzen
Die Proben wurden zur Verfügung gestellt von der BioProjekt-Datenbank des National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/bioproject/) und mit dem *Genome Analysis Toolkit* genotypisiert. Anschließend erfolgte das Mapping gegen das ARS-UCD1.2 Referenzgenom.

Probenname	BioProjekt Nummer	Geschlecht	Gbp [1]	Coverage	Anzahl gemappter Reads	Prozentualer Anteil gemappter Reads
ERR2694921	PRJEB27379	m	151,34	41,52	803.437.028	98,83
ERR2694941	PRJEB27379	m	151,43	37,74	747.353.123	99,32
ERR2694943	PRJEB27379	m	151,85	39,42	773.208.758	99,39
ERR2694946	PRJEB27379	m	151,60	39,11	777.252.594	99,40
ERR2694948	PRJEB27379	m	151,69	45,20	865.469.672	99,56
ERR2694950	PRJEB27379	m	154,10	40,83	802.882.089	99,41
ERR2694951	PRJEB27379	m	152,29	39,63	768.929.239	99,41
ERR2694953	PRJEB27379	m	152,00	41,23	807.612.066	99,43
ERR2694956	PRJEB27379	m	153,95	41,58	810.311.627	99,42
ERR2694957	PRJEB27379	m	154,82	40,46	788.967.830	99,43
ERR2694959	PRJEB27379	m	151,77	41,50	815.779.907	99,36
ERR2694960	PRJEB27379	m	155,37	38,46	756.543.714	86,40
ERR2694961	PRJEB27379	m	150,27	37,75	736.835.063	99,39
ERR2694976	PRJEB27379	m	103,69	62,30	1.220.468.674	99,50
SRR4449812	PRJNA350384	w	35,77	7,79	147.112.994	99,49

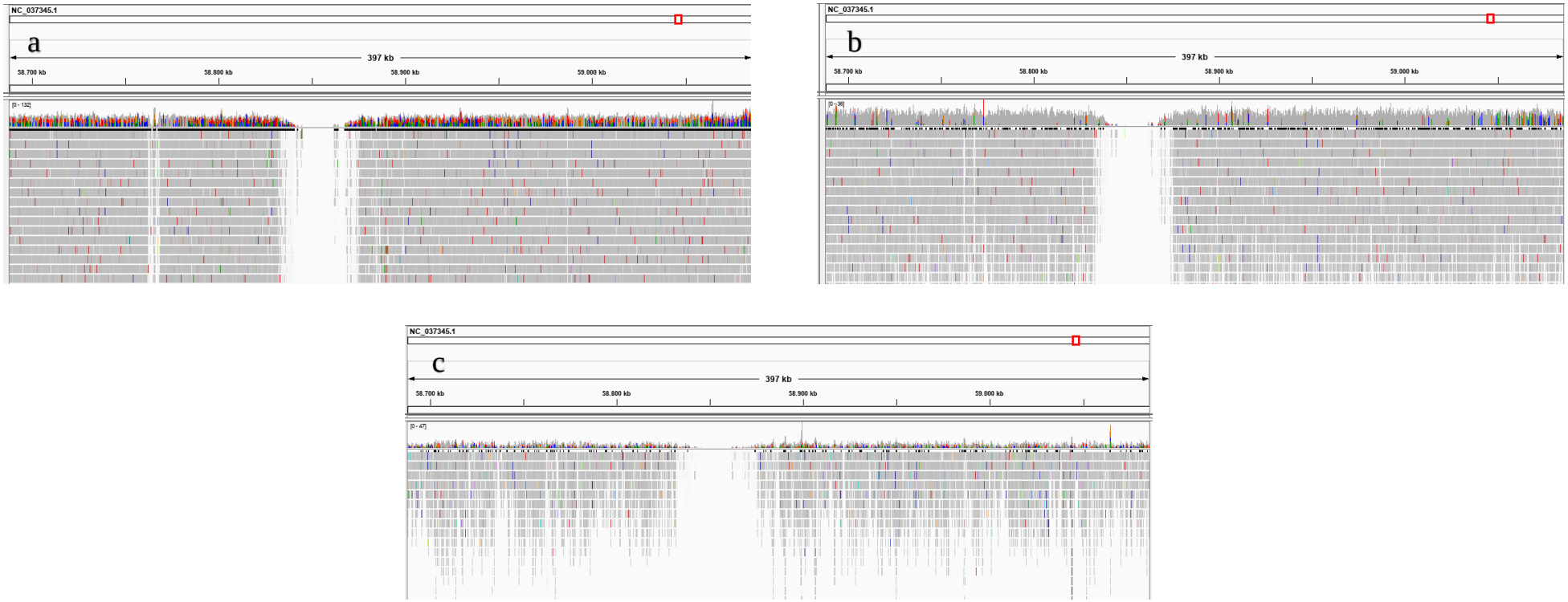
SRR4449830	PRJNA350384	m	28,92	9,59	182.305.950	99,05
SRR4450629	PRJNA350593	m	29,88	9,90	199.017.332	99,14
SRR4450630	PRJNA350593	m	17,83	5,53	116.947.212	99,09
SRR7461348	PRJNA477833	w	153,40	36,52	725.690.576	98,91
SRR7466787	PRJNA477833	m	151,35	8,31	159.816.156	99,03
SRR7466789	PRJNA477833	w	152,30	18,85	364.557.523	99,36

[1] Gesamtgröße der generierten Genomsequenz in Gigabasen

Anhang 2: IGV Darstellung der 16-Kb Lücke in den ONT sequenzierten Holstein Proben
Die Abbildungen zeigen jeweils den gleichen Bildausschnitt von 58.687.825 – 59.088.183 bp des Chromosom 18 von den Proben a) DHFnano01 (Gruppe q/Q); b) DHFnano02 (Gruppe q/Q); c) DHFnano03 (Gruppe Q/Q); d) DHFnano04 (Gruppe q/q) und e) DHFnano01/02. Alle Proben weisen den exakt identischen Bereich von ca. 16.000 bp auf, in welchem es nicht möglich war, DNA-Fragmente erfolgreich zu mappen. Diese 16-Kbp Lücke reicht in Abbildung a) von 58.846.131 – 58.861.709 bp; b) von 58.846.130 – 58.861.709 bp; c) 58.845.714 – 58.861.709 bp; d) 58.846.129 – 58.861.715 bp und e) 58.846.131 – 58.861.709 bp. Die Position des LRT-Peaks befindet sich bei 58.860.538 bp mit dem entsprechenden Konfidenzintervall von 58.343.346 – 59.432.662 bp.



Anhang 3: IGV Darstellung der 16-Kb Lücke in den ONT Proben der Rassen Kärntner Blondvieh und Fleckvieh
Der Bildausschnitt des BTA18 reicht von 58.687.825 – 59.088.183 bp. Sowohl das Kärntner Blondvieh (a) als auch das Fleckvieh (b) weisen die exakt gleiche 16-Kbp Lücke wie die Holsteinproben im Anhang 2 auf. In diesem Bereich zwischen a) 58.846.128 – 58.861.709 bp und b) 58.845.252 – 58.861.709 bp konnte kein DNA-Fragment gegen die Referenz gemappt werden.



Anhang 4: IGV Darstellung der 16-Kb Lücke in den Illumina sequenzierten Proben
Die Abbildungen a) Holstein ERR2694948; b) Hereford SRR8324584 und c) Brahman SRR2016745 stellen den Ausschnitt von 58.867.825 – 59.088.183 bp auf dem Chromosom 18 dar. Wie schon bei den ONT-Sequenzen (Anhang 2 und 3) ist auch hier deutlich die 16-Kbp Lücke zu sehen, die durch das vollständige Fehlen gemappter Reads hervorsticht. Dieser reicht in a) von 58.841.248 – 58.867.378 bp; b) 58.839.907 – 58.867.289 bp und c) 58.839.330 – 58.868.903 bp.

10. Danksagung

Zuallererst möchte ich der Tierzuchtforschung e.V. München für die Anstellung zur Durchführung dieser Arbeit danken. Insbesondere gilt mein Dank Herrn Prof. Dr. Georg Thaller und Herrn Dr. Ingolf Ruß, die die finanziellen Mittel dieses Forschungsvorhabens bereitstellten und mir den Besuch an der Vortragstagung der DGfZ und GfT ermöglichten.

Bei meinem Doktorvater PD Dr. Ivica Međugorac bedanke ich mich herzlichst für die Überlassung des sehr interessanten Themas, die Betreuung und Unterstützung in meiner Forschungszeit in der AG Populationsgenomik, sowie die kritische Durchsicht des Manuskripts.

Ein besonderer Dank gilt meiner Mitbetreuerin und zeitweilige Zimmerkollegin Dr. Elisabeth Hannemann für die Beantwortung meiner zahlreichen Fragen, Unterstützung auch während ihrer Elternzeit und die netten Gespräche in und außerhalb der Arbeit. Weiteres bedanke ich mich bei meinem zweiten Betreuer Ph.D. Maulik Upadhyay für die Hilfe in jeglichen bioinformatischen Fragen und Unterstützung in der Durchführung der Analysen dieses Forschungsprojekts.

Des Weiteren möchte ich mich bei Dr. Lilian Gehrke für die Unterstützung in diesem Projekt und ihr offenes Ohr für all meine Fragen trotz der Entfernung danken. Ein weiterer Dank gilt auch Frau Ottzen-Schirakow des Instituts für Tierzucht und Tierhaltung der Christian-Albrechts-Universität zu Kiel für die Bereitstellung der Proben.

Meinen zeitweiligen Mitdoktoranden Dr. Kim Eck und Dominik Lagler danke ich für die zwar kurze aber schöne Zeit, die unterhaltsamen Gespräche und die angenehme Arbeitsatmosphäre. Ein besonderer Dank gilt auch den Mitarbeitern der AG Populationsgenomik Renate Damian und Martin Dinkel, die mir durch so manche selbstgemachte Leckerei und unterstützenden Worte eine große Stütze während meiner Zeit waren. Auch danke ich allen weiteren Mitarbeitern des Instituts (Karina Schadt, Polyxeni Rizou und Jürgen Klawatsch) für die gemeinsame Zeit.

Außerdem möchte ich mich bei all meinen Freunden bedanken, die mich in meiner bisherigen Laufbahn begleiteten und mit denen ich zahlreiche lustige Stunden erleben durfte. Besonders danke ich Nik Reichl, Marie Schilloks,

Laura Müller sowie Margit und Vera Wunderlich für die unzähligen lustigen und vielleicht auch mal traurigen Momente, die wir bereits in unsere Studienzeit miteinander verbringen konnten. Ohne euch wäre meine Zeit hier in München nicht mal annährend so schön gewesen und ich möchte euch auch in Zukunft nicht mehr missen müssen.

Ein ganz besonderer Dank gilt meinem Freund Markus Pohl und seiner Familie. Danke an deine unendliche Geduld, deinen guten Zuspruch und, dass du immer an mich geglaubt hast. Ohne dich wäre diese Arbeit sicherlich nicht auf die gleiche Weise möglich gewesen. Ich freue mich auf die gemeinsame Zeit, die nun vor uns liegt.

Mein größter Dank gilt meiner Familie, die mich seit ich denken kann auf dem Weg zur Tierärztin unterstützt und mich in jeder schwierigen Phase meines Lebens weiter nach vorn gebracht hat. Ohne euch wäre ich nicht zu dem Menschen geworden der ich heute bin. Danke, dass ich mich immer auf euch verlassen kann, egal wie ausweglos die Situation auch scheint!