GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Veröffentlichung der Ergebnisse von Forschungsvorhaben im BMBF-Programm

AgroCluster: PHÄNOMICS – Ein systembiologischer Ansatz zur Genotyp-Phänotyp-Abbildung im Kontext von Leistung, Gesundheit und Wohlbefinden bei den Nutztieren Rind und Schwein – (Teilprojekt TP 1.3)

Förderkennzeichen: 0315536F

Zuwendungsempfänger: Universität zu Lübeck, 23562 Lübeck

Ausführende Stelle: Universität zu Lübeck –

Institut für Medizinische Biometrie und Statistik, 23562 Lübeck

Projektleitung: Herr Prof. Dr. Ziegler

Projektlaufzeit: 01.07.2010 bis 30.06.2014

Das diesem Bericht zugrundeliegende BMBF-Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 0315536F gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Inhaltsverzeichnis

I. Kurze Darstellung

1. Aufgabenstellung

Genomweite Assoziationsstudien im Menschen haben seit 2007 maßgeblich zur Identifikation vieler Assoziationen zwischen Einzelnukleotidpolymorphismen (engl: single nucleotide polymorphisms, kurz: SNPs) und komplexen genetischen Erkrankungen beigetragen. Die funktionelle Bedeutung muss allerdings für die meisten dieser Loci noch geklärt werden.

Ein integrativer genomischer Ansatz, der dem zentralen Dogma der Molekularbiologie folgt, erscheint vielversprechend. Hiermit könnte es möglich sein den Pfad von der DNA über die RNA über das Proteom sowie dem Metabolom zu intermediären Phänotypen und abschließend zu Krankheitsphänotypen abzubilden. Mit relativ großen Messfehlern sind 2DE Proteindaten behaftet. Die derzeit verfügbaren Metabolom-Chips sind in ihrer Größe erheblich limitiert.

Daher werden in den nächsten Jahren SNP-Genexpressions-Assoziationen – sowohl beim Menschen als auch bei Nutztieren - von großer Bedeutung sein. Zum einen sind beide Technologien im Hochdurchsatzbereich gut etabliert. Zum anderen werden Expressionen, d.h. Transkripte, direkt durch Variationen auf genomischer Ebene beeinflusst. Darüber hinaus ist der Vorteil der SNPs, dass diese keinen Zufallsschwankungen oder Modifikationen unterliegen und Ausgangspunkt weiterer molekularer Biomarker sind.

In SNP-Expressions-Assoziationsstudien, die auch als eQTL-Studien (engl. expression Quantitative Trait Loci) bezeichnet werden, wird das Expressionsniveau als quantitatives phänotypisches Merkmal aufgefasst und auf Assoziationen mit genetischen Markern hin untersucht. Es sollen also zum Beispiel SNPs identifiziert werden, welche das Expressionsniveau eines Gens beeinflussen.

Es werden Genotypen von genomweit verteilten SNPs und die Expressionswerte vieler bekannter Transkripte in relevantem Gewebe bestimmt. Nach einer Qualitätskontrolle liegen für jedes Individuum diskrete Genotypdaten (z.B. AA, AG, GG bzw. als 0, 1, 2 kodiert) für die SNPs und quantitative (logarithmierte) Expressionsdaten für die Transkripte vor. Für jede mögliche Kombination von SNPs und Transkripten wird ein statistischer Test auf Assoziation durchgeführt. Wenn sich die mittleren Expressionswerte eines Gens bei Trägern verschiedener Genotypen nicht nennenswert voneinander unterscheiden, ist die Expression offenbar unabhängig vom Genotyp, es liegt keine Assoziation vor. Falls sie sich hingegen voneinander unterscheiden, liegt eine Assoziation vor. Es muss also ein statistischer Test auf einen Lageunterschied bei drei Stichproben durchgeführt werden.

Vor Projektbeginn wurde gezeigt, dass die Varianzanalyse, die als Standardverfahren zur SNP-Expressions-Assoziationsanalyse eingesetzt wird, in vielen Situationen zu liberalen Ergebnissen führt [1]. Insbesondere ist dieses bei schiefen Verteilungen oder Verteilungen mit Ausreißern der Fall. Bei Daten von Expressions-Niveaus liegt jedoch trotz aufwändiger Vorverarbeitung häufig keine Normalverteilung der Daten vor. Auch andere Standardverfahren wie der Kruskal-Wallis-Test könnte in diesen Fällen eine geringere statistische Macht haben als andere statistische Tests, die speziell für den Fall z.B. rechtsschiefer Daten entwickelt wurden. Deshalb ist davon auszugehen, dass viele statistisch signifikante Ergebnisse solcher Standardverfahren tatsächlich falsch positive Ergebnisse sind, andererseits einige wahre Assoziationen nicht als solche identifiziert werden können.

Ziel des Teilprojekts ist es daher, publizierte statistische Verfahren zur Analyse von SNP-Expressions-Assoziationsanalysen zu identifizieren und sie auf ihre Gültigkeit und praktische Anwendbarkeit in der Hochdurchsatzsituation zu überprüfen. Darüber hinaus soll ein neues statistisches Verfahren entwickelt werden, welches eine größere statistische Macht besitzen. Hier bieten sich insbesondere adaptive statistische Verfahren an, bei denen zunächst ein spezifisches Testverfahren unter Verwendung einer Selektorstatistik – nicht eines statistischen Vorschalttests – ausgewählt wird.

Diese Verfahren müssen im Rahmen des Teilprojektes in Standard-Software effizient implementiert werden, um in Hochdurchsatzstudien auch tatsächlich anwendbar zu sein. Auch soll deren Verhalten in spezifischen Verteilungssituationen mittels Monte-Carlo Simulationen untersucht werden. Schließlich soll das bevorzugte Verfahren im Rahmen des Projektverbunds in substanzwissenschaftlichen Projekten genutzt werden, um neue funktionelle Variationen und Biosignaturen zu identifizieren und zu validieren.

2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Der Antragsteller verfügt über langjährige Erfahrungen im Bereich der statistischen Analyse von Hochdurchsatzdaten vom Menschen, sowie mit dem Gebrauch von Simulationsstudien zum Vergleich von statistischen Testverfahren. Ein wichtiger Teil der oben genannten Ziele konnten dennoch nur durch die Zusammenarbeit mit anderen Projektgruppen erfolgreich durchgeführt werden. Besonders die praktische Erfahrung des Leibniz-Instituts für Nutztierbiologie FBN Dummerstorf mit Daten von Nutztieren aus mehreren molekularen Ebenen war dabei sehr hilfreich. Die Struktur als Verbund

ermöglichte insbesondere eine zielgerichtete Zusammenarbeit mit anderen Teilprojektgruppen bei der Vorverarbeitung und den Analysen realer Daten einer speziellen Rinderpopulation.

3. Planung und Ablauf des Vorhabens

Die Planung und Durchführung der Untersuchungen gliederte sich entsprechend des Projektantrages in die unten genannten Arbeitspakete. Die Arbeit des Antragstellers konnte wie geplant im Juli 2010 begonnen werden. Bedingt durch einen verzögerten Erhalt von SNP- und Expressionsdaten konnte die statistische Analyse (AP4 und AP5) erst mit mehrmonatiger Verzögerung begonnen werden. Insgesamt konnten dank effizienter Implementierung der Tests alle Arbeitspakete wie geplant im Juni 2014 abgeschlossen werden.

AP1 Literaturrecherche zu adaptiven Verfahren und Entscheidung für geeignete Ansätze für SNP-Expressions-Assoziationsstudien

AP2 Implementierung eines oder mehrerer adaptiver Verfahren für SNP-Expressions-Assoziationsstudien in Standard-Software

AP3 Evaluierung der adaptiven Verfahren hinsichtlich Fehlerniveaus und Güte

AP4 Festlegung der Gewichte für genomweite SNP/Transkriptom-Analysen aus „omics"-Netzwerkstrukturen

AP5 Anwendung in praxisrelevanten Datensätzen zur Aufklärung der molekularen Basis identifizierter Biosignaturen

4. Wissenschaftlicher Stand

Zu Beginn wurde gemäß AP1 eine systematische Literaturrecherche zum Thema „Adaptive Verfahren" durchgeführt. Ziel dieser Literaturrecherche und der nächsten weiterführenden Arbeitspakete dieses Teilprojektes war es, ein robustes statistisches Testverfahren für das Lageproblem bei drei Stichproben zu finden. Dieses Verfahren sollte auch bei nicht-normalverteilten Daten, die häufig in SNP-Expressions-Assoziationsstudien auftreten, möglichst wenige falsch-positive Ergebnisse liefern, aber gleichzeitig möglichst viele wahre Assoziationen identifizieren. Als besonders erfolgsversprechend gelten ein nicht-parametrischer adaptiver Test von Büning [2] und dessen Modifikation von Beier [3], die beide auf einem Ansatz von Hogg, Fisher und Randles [4] basieren. Als weitere interessante Ansätze wurden ein ebenfalls nicht-parametrisches adaptives Verfahren von O'Gorman [5] und ein parametrisches Verfahren von Keselman [6] identifiziert. Letzteres basiert auf einer bekannten Modifikation der Varianzanalyse nach Welch [7], bei dem die Parameter adaptiv getrimmt werden. Diese vier statistischen Verfahren bildeten die Grundlage für die Arbeiten an diesem Teilprojekt. Ein Vergleich dieser adaptiven statistischen Testverfahren war in der Literatur nicht zu finden. Auch konnte keine Anwendung derartiger Tests im Rahmen von SNP-Expressions-Assoziationsstudien in der Fachliteratur gefunden werden.

5. Zusammenarbeit mit anderen Stellen.

Dieses Teilprojekt (TP1.3) wurde im Rahmen des Kompetenznetzes PHÄNOMICS als Teil des Verbundprojektes VP1 - Integrative Bioinformatik - durchgeführt. Gegenstand dieses Verbundprojektes ist der Aufbau einer „Animal Trait Ontology" (ATO)-Plattform, die Bereitstellung einer zentralen Datenbank für das PHÄNOMICS-Konsortiums, welche die gegenseitige Information

und Datenverfügbarkeit sichert, (TP1.1) und die systembiologische Modellierung der Genotyp-Phänotyp-Abbildung (TP1.2, TP1.3, TP1.4).

Der Bioinformatik-Ansatz zur systembiologischen Modellierung der Genotyp-Phänotyp-Abbildung kombiniert Kontextwissen mit Hochdurchsatzdaten und gemessenen genotypisch/phänotypischen Parametern. Die im Verbundprojekt VP3 ermittelten 'omics'-Daten bilden dabei zusammen mit den erfassten Leistungs- und Gesundheitsmerkmalen sowie den im Verbundprojekt VP2 entwickelten und gemessenen Indikatoren des Wohlbefindens die Datengrundlage.

Im Rahmen des Kompetenznetzes wird durch die ATO ein wichtiger Beitrag zur besseren wissenschaftlichen Begründung des Merkmalskomplexes Wohlbefinden erwartet. Das durch diese Arbeit entwickelte statistische Verfahren zur Identifizierung von Assoziationen zwischen SNPs und Transkripten funktionaler Gene oder auch Proteinmesswerten stellt dabei ein Bindeglied zur Aufklärung des molekularen Zusammenhangs zwischen genetischer Variation und messbaren quantitativen Phänotypen wie z.B. der Milchleistung oder auch Indikatoren für das Wohlbefinden von Nutztieren dar.

Sowohl bei der Vorverarbeitung, als auch bei den Analysen der Realdaten fand mit den Teilprojekten TP1.4 und TP3.2 – aus dem Leibniz-Institut für Nutztierbiologie FBN Dummerstorf - ein reger Austausch von Kenntnissen, Daten und Programmcode im Rahmen von Projekttreffen und zahlreichen Telefonaten und Emails statt.

II. Eingehende Darstellung

1. Ergebnisse

Die Zuwendungen wurden entsprechend der Zielsetzung im Projektantrag eingesetzt. Im Folgenden sollen am Beispiel einzelner Arbeitspakete die Verwendung der Zuwendungen dargelegt werden.

AP1

Gemäß Arbeitspaket AP1 wurde zu Beginn des Teilprojektes eine systematische Literaturrecherche durchgeführt. Ziel dieser Literaturrecherche und der nächsten weiterführenden Arbeitspakete dieses Teilprojektes ist es, ein robustes statistisches Testverfahren für das Lageproblem bei drei Stichproben zu finden. Dieses Verfahren sollte auch bei nicht-normalverteilten Daten, die häufig bei Expressionsdaten auftreten, möglichst wenige falsch-positive Ergebnisse liefern, aber gleichzeitig möglichst viele wahre Assoziationen identifizieren. Die Herausforderung dabei ist, aus der Vielzahl von theoretisch anwendbaren Tests einen für die vorliegenden Daten geeigneten auszuwählen. In Anbetracht der Tatsache, dass Expressionsdaten ganz unterschiedlichen Verteilungsformen unterliegen können, ist es von vornherein aussichtslos, nach einem einzigen Test zu suchen, der für alle möglicherweise auftretenden Fälle optimal geeignet ist. Eine Lösung dieses Problems bieten adaptive Verfahren, welche, abhängig von den Eigenschaften der vorliegenden Daten, jeweils einen geeigneten Test auswählen.

Die Recherche war sehr zeitintensiv, da eine Vielzahl von unterschiedlichen Ansätzen in der Literatur zu finden ist und zahlreiche Veröffentlichungen zur genaueren Einordnung gründlich gelesen werden mussten um die Praxis-Tauglichkeit in groben Zügen bewerten zu können. Zunächst wurden verschiedene Ansätze kategorisiert (z.B. die Art der Selektorstatistiken), im Hinblick auf die Anwendbarkeit für oben beschriebenes Testproblem bewertet und miteinander verglichen.

Vielversprechend erscheinende Ansätze wurden identifiziert, welche in den nächsten Arbeitsschritten zunächst implementiert und dann mit Hilfe von Simulationsstudien auf Eigenschaften wie Güte und Robustheit hin untersucht werden sollen, insbesondere für den Fall nicht-normalverteilter Daten.

Für die Suche wurden drei elektronische Datenbanken mit spezifischen Anfragen durchsucht. Verwendet wurden die Datenbanken Zentralblatt MATH (1826 bis heute), Web of Science (1945 bis heute) und Current Index to Statistics (1975 bis heute). Die letzte Suchanfrage fand am 10.03.2011 statt.

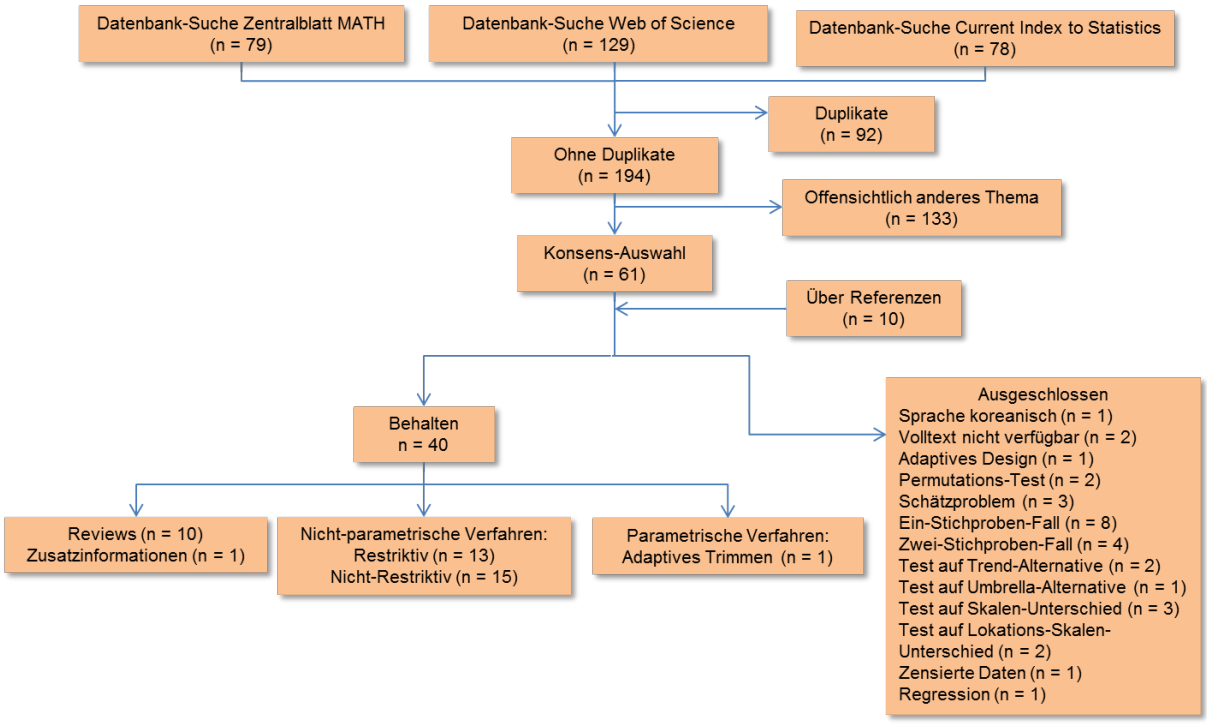

Abbildung 1: Flussdiagramm zur Literaturauswahl

Wie aus dem Flussdiagramm in Abbildung 1 hervorgeht wurden durch die Such-Anfragen in den drei oben genannten Datenbanken insgesamt 286 Treffer erzielt, von denen allerdings 92 Duplikate waren. Es blieben also nach der ersten Filterung 194 Veröffentlichungen übrig, die manuell betrachtet werden mussten. Bei der Betrachtung der Zusammenfassungen konnten 133 themenfremde Veröffentlichungen herausgefiltert werden, so dass noch 61 potentiell interessante übrig blieben. Da über Referenzen noch weitere zehn potentiell nützliche Artikel dazukamen mussten insgesamt 71 Artikel gelesen werden, um sie in eine der Kategorien einzuordnen.

Insgesamt wurden 40 Artikel als interessant eingestuft und intensiver betrachtet. Eine Reihe davon sind Review-Artikel (z.B. [8], [9], [10], [11], [12]), die dem interessierten Leser einen guten Überblick über die verschiedenen Ansätze geben, und zum anderen zahlreiche methodische Artikel, in denen oft ähnliche und überwiegend abgewandelte Ansätze beschrieben werden.

Das einzige hierbei identifizierte parametrische Verfahren von Keselman [6] basiert auf einer Modifikation der Varianzanalyse nach Welch [7]. Als robuster Lageparameter wird hier der getrimmte Mittelwert verwendet, wobei adaptiv, d.h. abhängig von den Verteilungen, der

Trimmanteil auf die rechte und linke Seite der Verteilungen aufgeteilt wird. Dieser Ansatz erscheint vielversprechend. Die nicht-parametrischen Ansätze lassen sich laut Hušková [11] in zwei Gruppen einteilen: Restriktive und Nicht-Restriktive Verfahren.

Restriktive beziehungsweise partiell adaptive Verfahren schätzen mit Hilfe von sogenannten Selektorstatistiken grob die Verteilungsform der Daten und wählen aus einer zuvor festgelegten Familie von Tests den für die beobachtete Situation am besten geeigneten Test aus. Der bekannteste Test dieser Art für den 2-Stichprobenfall von Hogg, Fisher und Randles [4] wurde von Büning [2] für den Fall von c Stichproben erweitert. Diese adaptiven Tests nutzen rangbasierte Selektorstatistiken, welche die Schiefe und die Länge der Verteilungsschwänze schätzen, sowie eine Reihe von nicht-parametrischen linearen Rang-Tests. Es gibt zahlreiche sehr ähnliche Abwandlungen dieser Verfahren (z.B. [13], [14]), darunter auch Permutationstests (z.B. [15]), die wir jedoch außer Acht lassen, da solche rechenintensiven Verfahren in Anbetracht der Menge an durchzuführenden Tests praktisch nicht durchführbar sind. Die Methode aus [2] erscheint leicht auf die oben beschriebene Problemstellung übertragbar zu sein. Vor allem ist dieses Verfahren interessant, da es – je nach Wahl der Selektorstatistiken – verteilungsfrei oder immerhin asymptotisch verteilungsfrei ist, also das vorgegebene Signifikanzniveau auch bei nicht-normalverteilten Daten einhält. Daher soll dieses Verfahren in den nächsten Arbeitsschritten vorrangig untersucht werden. Es gibt für den 2-Stichprobenfall auch Ansätze, welche auf U-Statistiken basieren ([16], [17]). Der Aufwand für die Umsetzung auf den 3-Stichprobenfall erscheint allerdings in Anbetracht des geringen erhofften Zugewinns an Güte und Robustheit fraglich, deshalb werden wir uns vorerst auf andere Ansätze konzentrieren.

Nicht-Restriktive beziehungsweise voll adaptive Verfahren unterscheiden sich dahingehend, dass sie Informationen aus den vorliegenden Daten in die Gewichte der Teststatistik mit einfließen lassen und diese somit kontinuierlich an die Daten anpassen können. Der erste interessante aber sehr komplizierte Ansatz dieser Art für den Fall von zwei Stichproben wird in [18] beschrieben. Wesentlich einfacher verständlich ist das Verfahren von O'Gorman [19] bzw. [5], welches auch für den c-Stichprobenfall modifiziert wurde [5]. Letzteres erscheint ebenfalls sehr interessant und soll nach Möglichkeit in Zukunft implementiert und weiter untersucht werden.

AP2

Gemäß Arbeitspaket AP2 wurden die vier adaptiven Testverfahren von Büning [2], Beier [3], O'Gorman [5] und Keselman [6] in der Statistiksoftware R implementiert.

Die Auswahlkriterien der Tests von Büning und Beier wurden mittels einer sehr zeit- und rechenaufwendigen Monte-Carlo Simulationsstudie im Hinblick auf die Anforderungen von SNP-Expressions-Assoziationsstudien überprüft und ein modifiziert Test entwickelt. Beide Verfahren benutzt Informationen der Datenverteilung, um für die vorliegenden Daten einen geeigneten Test auszuwählen. In einem ersten Schritt werden die Schiefe und die Länge der Verteilungsschwänze geschätzt. Auf Grundlage dieser quantil-basierten Selektorstatistiken wird anschließend eine lineare Rangteststatistik aus einer Schar linearer Rangstatistiken für das statistische Testen ausgewählt. Diese Strategie ist grundlegend verschieden von dem kritisch zu beurteilenden Vorgehen, einen Vorschalttest zu nutzen, um auf der Basis dieses statistischen Tests den eigentlichen statistischen Test von Interesse anzuschließen. Denn der Ansatz mit einem p-Wert-basierten Vorschalttest führt zu einem erhöhten Typ-I-Fehler [20], wohingegen ein adaptiver Test, der rangbasierte Selektorstatistiken mit verteilungsfreien Tests kombiniert, das nominale Signifikanzniveau einhält [8].

Die Berechnung der Selektorstatistiken erfolgt bei beiden Verfahren in gleicher Weise, bei der nachgeschalteten Auswahl der linearen Rangstatistiken unterscheiden sich die Verfahren jedoch. Die Entscheidungsregeln beider Verfahren sind in Abbildung 2 dargestellt. Ein Vergleich der beiden Methoden war in der Literatur nicht zu finden. Wir haben dies nun in einer aufwendigen Monte-Carlo Simulationsstudie für eine Vielzahl von Verteilungen mit variierender Schiefe und Länge der Verteilungsschwänze durchgeführt. Dabei wurde ein neuer adaptiver Test entwickelt, der die Vorteile beider oben angeführten Verfahren in sich vereinigt und deshalb in speziellen Situationen eine größere statistische Macht hat, ohne dabei an Robustheit zu verlieren.

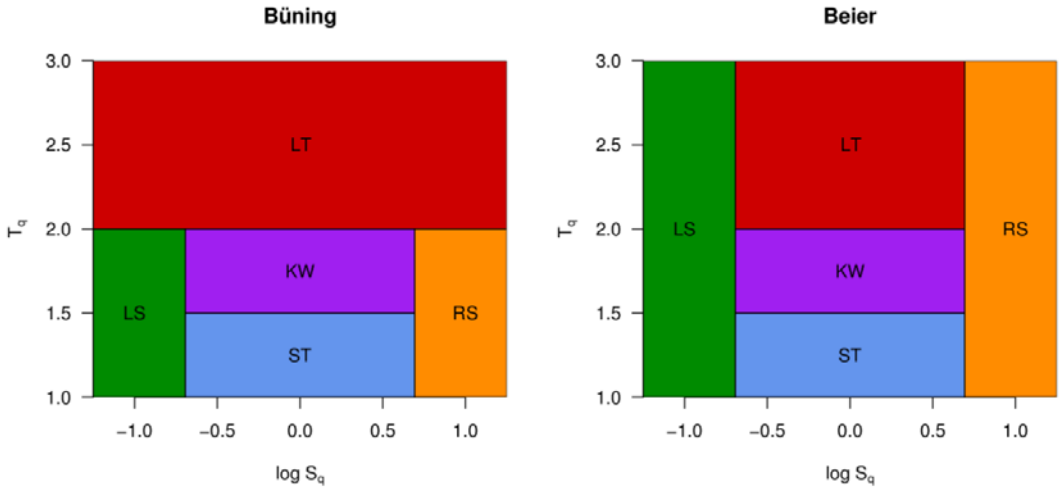

Abbildung 2: Entscheidungsregeln nach Büning [2] und Beier [3]. Für jedes Wertepaar log(S_q) und T_q – den Punktschätzern für die Schiefe bzw. die Länge der Verteilungsschwänze - ist die entsprechende anzuwendende lineare Rangteststatistik angegeben (KW: Kruskal-Wallis Test, LS: Linksschiefe Test, LT: Long Tails test, RS: Rechtsschiefe Test, ST: Short Tails Test). Diese Abbildung stammt aus der eingereichten Veröffentlichung. Der ST ist in der Literatur unter der Bezeichnung Gastwirth Test bekannt und der RS als Hogg-Fisher-Randles Test.

Für die Monte-Carlo Simulationsstudie wurden dieselben Schätzer für die Schiefe (S_q) und die Länge der Verteilungsschwänze (T_q) genutzt wie von Büning und Beier – wobei die Darstellung mit logarithmiertem Schiefeschätzer aus symmetriegründen vorzuziehen ist. Zusätzlich zu den linearen Rangstatistiken wurde der Median Test (MED) betrachtet, der ebenfalls als lineare Rangstatistik betrachtet werden kann.

Für die Simulation der Daten wurde die bislang nur wenig bekannte g-und-k-Verteilungsfamilie verwendet [21]. Mit Hilfe dieser Verteilungsfamilie ist es möglich, Verteilungen mit spezifischen vorab gewählten Schiefen und Schwanzlängen zu simulieren. Dies erwies sich als außerordentlich hilfreich, da für ganz verschiedene Verteilungsformen, die wiederum je einem Punkt auf der S_q-T_q-Entscheidungsebene entsprechen, eine lineare Rangteststatistik mit möglichst optimalen Eigenschaften identifiziert werden sollte, um so den adaptiven Test zu optimieren. Die theoretischen Selektorstatistiken der auf diese Weise simulierten Szenarien, die ein Gitter in der log S_q-T_q-Ebene bilden, sind in Abbildung 3 als graue Punkte dargestellt.

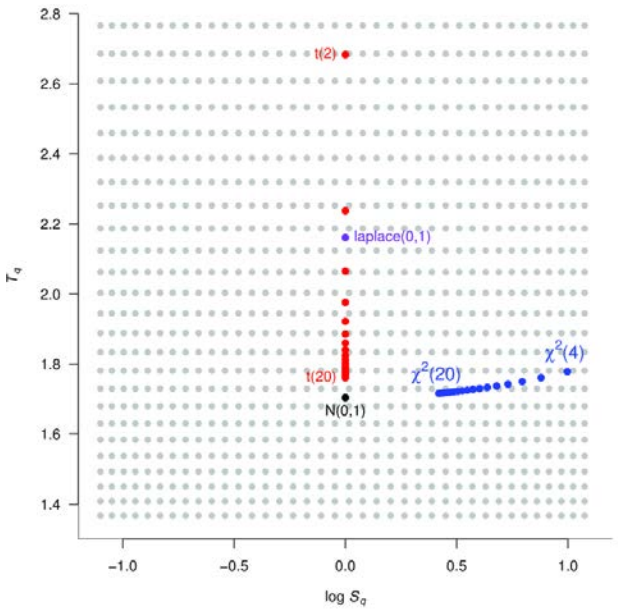

Abbildung 3: Entscheidungsebene der Selektorstatistiken log(S_q) und T_q. Die grauen Punkte in dieser Abbildung entsprechend den Punktschätzern der simulierten Szenarien, die mit der g-und-k-Verteilungsfamilie erzeugt wurden. Zum Vergleich wurden die entsprechenden Schätzer bekannter Verteilungen abgebildet.

Für jedes Szenario wurden die Daten mit einer Gesamtfallzahl von 600 in drei Gruppen gemäß einem Hardy-Weinberg-Gleichgewicht mit variierender Häufigkeit des selteneren Allels (minor allele frequency; MAF) erzeugt. Dabei waren die Verteilungsformen in den drei Gruppen prinzipiell jeweils gleich. Um sowohl die Güte als auch die Robustheit der linearen Rangstatistiken überprüfen zu können, wurden alle Szenarien sowohl ohne als auch mit einem theoretischen Unterschied in den Medianen in den drei Genotyp-Gruppen 1000× repliziert. Ein Test sollte hier als robust gelten, wenn der geschätzte Typ-I-Fehler gemäß des liberalen Bradley-Kriteriums [22] zwischen $0{,}5\ \alpha$ und $1{,}5\ \alpha$ liegt, wobei $\alpha = 0{,}05$ das gewählte Signifikanzniveau des Tests ist.

Die empirische Güte von RS und LS sind wie erwartet für rechts- bzw. linksschiefe Verteilungen am größten, die des ST ist bei kurzschwänzigen Verteilungen am größten. Der KW hat eine ähnliche empirische Güte wie ST, ist aber für Verteilungen mit mittleren Schwänzen optimal. MED hat für langschwänzige Verteilungen die höchste empirische Güte, in allen anderen Fällen dafür eine sehr geringe. Die empirische Güte für LT ist bei allen Verteilungsformen relativ ähnlich.

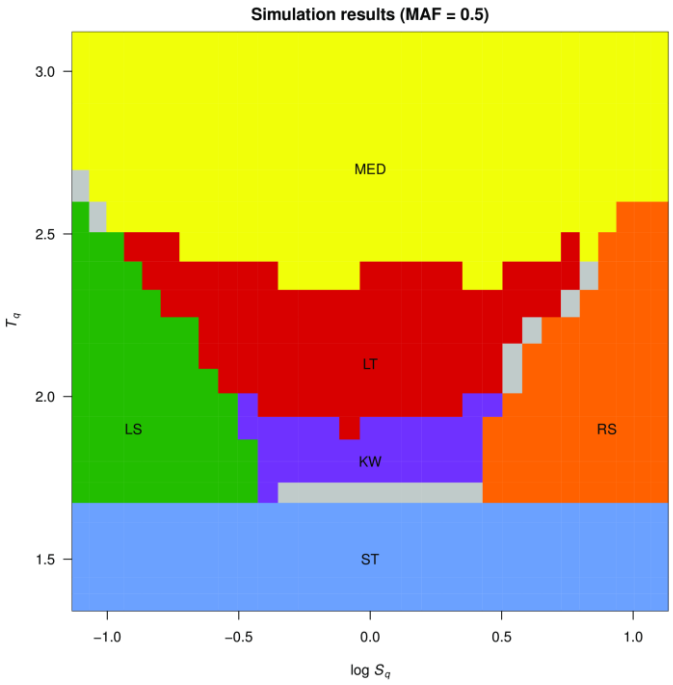

Abbildung 4: Optimale lineare Rangteststatistiken für simulierte Szenarien unter der Alternativhypothese bei einer MAF von 0,5. Für jedes Wertepaar log(S_q) und T_q wurde die empirische Güte anhand der Simulationsszenarien interpoliert. In den grauen Bereichen sind mehr als ein Test optimal. (KW: Kruskal-Wallis Test, LS: Linksschiefe Test, LT: Long Tails test, RS: Rechtsschiefe Test, ST: Short Tails Test, MED: Median Test). Diese Abbildung stammt aus der eingereichten Veröffentlichung.

Abbildung 4 zeigt für jedes simulierte Szenario die lineare Rangteststatistik mit der größten empirischen Güte. Auf Grundlage dieser Abbildung können neue Entscheidungsgrenzen für ein adaptives Verfahren festgelegt werden, so dass für bestimmte Verteilungsformen jeweils die optimale lineare Rangstatistik ausgewählt wird.

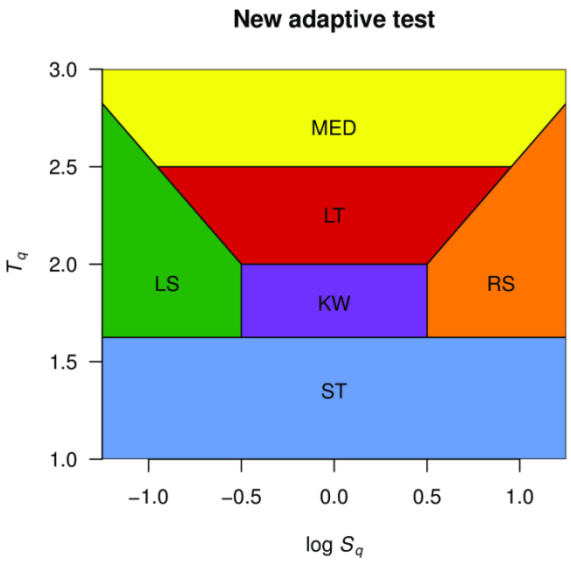

Abbildung 5: Neues Entscheidungsschema für einen adaptiven Test. Auf Grundlage der empirischen Güte in der Simulationsstudie wurden die Grenzen so festgelegt, dass für bestimmte Verteilungsformen jeweils die optimale lineare Rangstatistik ausgewählt wird. (KW: Kruskal-Wallis Test, LS: Linksschiefe Test, LT: Long Tails test, RS: Rechtsschiefe Test, ST: Short Tails Test, MED: Median Test). Diese Abbildung stammt aus der Veröffentlichung [23].

Abbildung 5 zeigt das neue Entscheidungsschema, das auf Grundlage der empirischen Güte in der Simulationsstudie angepasste wurde. Im Vergleich mit den bisherigen Verfahren von Büning und Beier fällt zunächst auf, dass MED als lineare Rangstatistik in die Auswahl mit aufgenommen wurde, da dieser auf den simulierten Daten mit langen Verteilungsschwänzen die größte empirische Güte hatte. Als weitere wesentliche Änderungen seien zu nennen, dass ST im Bereich für kurze Verteilungsschwänze nun auch bei schiefen Verteilungen ausgewählt wird, und dass die Entscheidungsgrenzen bei langschwänzigen, schiefen Verteilungen schräg verlaufen, was in diesem Bereich einem Kompromiss zwischen den Verfahren von Büning und Beier zu sein scheint.

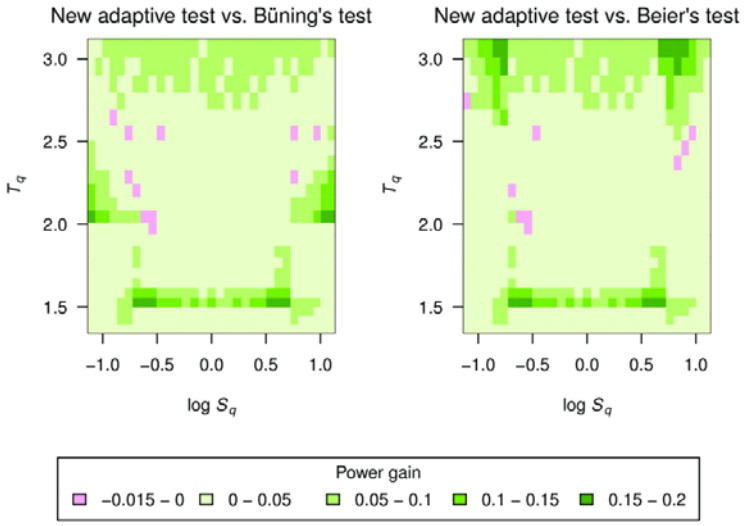

Abbildung 6: Heatmap des Güte-Vergleichs zwischen dem neuen adaptiven Test und den Tests von Büning bzw. Beier. Für jedes Wertepaar log(S_q) und T_q wurden die Differenzen der empirischen Güte des neuen adaptiven Tests und dem adaptiven Test von Büning bzw. Beier anhand der Simulationsergebnissen interpoliert. An den Stellen, wo die Entscheidungsgrenzen geändert wurden, konnte ein Güte-Gewinn von bis zu 20% erzielt werden. Diese Abbildung stammt aus der Veröffentlichung [23].

Gerade in diesen Bereichen gibt es bei dem so entwickelten neuen adaptiven Test auch den größten Gewinn an Güte, was der direkte Vergleich zwischen dem neuen adaptiven Test und den Tests von Büning bzw. Beier in Abbildung 6 deutlich macht. An den Stellen, wo die Entscheidungsgrenzen geändert wurden, konnte ein Güte-Gewinn von bis zu 20% erzielt werden. Der Güte-Verlust an anderen Stellen ist mit maximal 1,5% vernachlässigbar klein.

Es konnte also wie geplant mit Hilfe der oben beschriebenen Simulationsstudie ein neuer adaptiver Test entwickelt werden, der sowohl robust ist, als auch über eine größere statistische Macht bei nicht-normalverteilten Daten verfügt als das adaptive Verfahren von Büning und Beier. Dieser neue statistische Test (im weiteren „Test von Szymczak“ genannt) konnte in Statistics in Medicine [23] veröffentlicht werden.

AP3

Für die Evaluierung des neuen Tests von Szymczak et al. [23] hinsichtlich Fehlerniveaus und Güte und dem Vergleich mit den adaptiven Tests von O’Gorman [5] und Keselman [6] sowie den üblichen Standard-Verfahren Varianzanalyse und Kruskal-Wallis-Test wurden gemäß AP3 weitere Simulationsstudien durchgeführt, bei denen die Grundkonzepte der Simulationsstudie aus AP2 genutzt werden konnten. Für diese Untersuchungen genügte es, Daten für zwei Gruppen zu simulieren, da die Eigenschaften der Tests so deutlicher zu erkennen sind, als bei drei Gruppen. Daraus gewonnene Erkenntnisse sind auf die Situation von drei oder mehr Gruppen übertragbar.

Auch in den vier Simulationsstudien aus diesem Arbeitspaket wurden Daten mittels der g-und-k-Verteilungen mit verschiedenen Schiefen und Verteilungsschwanzlängen simuliert. Die Eigenschaften der Tests wurden in Simulationsstudie I im reinen Shift-Modell, also mit homogenen Varianzen und in Simulationsstudie II im Model mit Shift und inhomogenen Varianzen analysiert.

In Simulationsstudie I hielten alle untersuchten Tests das Typ I Fehlerniveau ein. Für relativ normalverteilte Daten hatten die drei adaptiven Tests eine große statistische Macht, ebenso die beiden Standardtests. Doch im Fall von sehr schiefen und sehr langsschwänzigen Verteilungen hatten nur noch die Tests von Szymczak und O’Gorman eine große Macht.

In Simulationsstudie II hielten die meisten Tests in vielen simulierten Szenarios das vorgegebene Signifikanzniveau nicht ein. Dabei wurden die Abweichungen im Allgemeinen größer, je weniger normalverteilt die Daten waren, vor allem bei sehr schiefen Daten. Der U-Test, welcher stellvertretend für den Kruskall-Wallis-Test untersucht wurde, erwies sich bei den hier simulierten Daten als recht robust. Wären die Unterschiede in den Varianzen zwischen den Gruppen jedoch größer gewesen, so hätte auch dieser Test einen deutlich erhöhten Typ I-Fehler gezeigt, wie aus früheren Arbeiten, z.B. [1], hervorgeht.

Insgesamt erwies sich der parametrische adaptive Test von Keselman [6] in beiden Simulationsstudien als robuster und machtvoller als der t-test. Doch im Fall nicht-normalverteilter Daten sind diese beiden Tests deutlich weniger robust und weniger machtvoll als die untersuchten nicht-parametrischen Tests.

Die Ergebnisse der beiden nicht-parametrischen Tests von Szymczak [23] und O’Gorman [5]waren relativ ähnlich. In Simulationsstudie II wurde deutlich, dass beide Tests bei schiefen Verteilungen Probleme mit der Robustheit haben. Dafür hatten beide Tests in fast allen simulierten Situationen in Simulationsstudie I die größte statistische Macht von allen untersuchten Tests, wobei der Test von

Szymczak bei Daten mit sehr langen Verteilungsschwänzen sogar die größte Güte von allen untersuchten Tests hatte.

Die Ergebnisse dieser Simulationsstudien konnten erfolgreich im Journal Methods of Information in Medicine publiziert werden [24]. Zusammenfassend lässt sich sagen, dass der in AP2 entwickelte neue adaptive Test zwar auch nicht in allen Situationen robust ist, doch im Vergleich mit den anderen untersuchten Verfahren hat er sehr gut abgeschnitten und hatte sogar in vielen simulierten Situationen die größte statistische Macht. Dieser Test scheint also eine ausgezeichnete Alternative zu den üblichen Standardmethoden und wurde deshalb auch wie geplant in den weiteren Arbeitspaketen zur Analyse von molekularen Daten verwendet.

Ein weiteres zentrales Ergebnis dieser ersten drei Arbeitspakete ist die Entwicklung des Software-Paketes zur Berechnung des adaptiven Tests in der Programmiersprache R. Die Software ist einfach zu bedienen und flexibel einsetzbar, da dieser auch bei Gruppenvergleichen von Messungen anderer quantitativer Merkmale angewendet werden kann. Der Programmcode wurde auch anderen Teilprojekten des Projektverbunds zur Verfügung gestellt, die explizit Interesse an der Anwendung der Methode auf andere Datensätze zu Ausdruck gebracht haben. Denn bei vielen Messungen im Rahmen des Projektverbunds, z.B. von quantitativen Verhaltensmerkmalen, sind die Daten überwiegend nicht normalverteilt.

AP4

Im Rahmen von AP4 sollte eine Assoziationsanalyse zwischen SNPs und Transkripten bei einer speziellen Rinderpopulation durchgeführt werden. Gegenstand der Analysen waren Daten von 145 Kühen aus F2-Familien, die aus einer Kreuzung zwischen Deutschen Holstein-Kühen (Milchkuh) und Charolais-Bullen (Fleischrind) generiert wurden [25]. Das Kreuzungsschema ist in Abbildung 7 dargestellt.

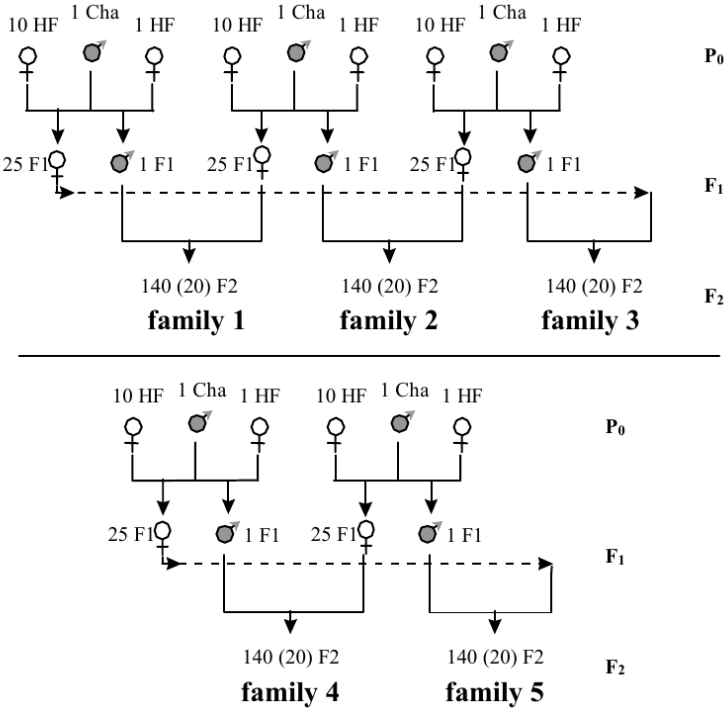

Abbildung 7: Geplantes Kreuzungsschema zwischen Deutschen Holstein-Kühen und Charolais-Bullen. Daten von Kühen aus der F2-Generation waren Gegenstand unserer Analysen in AP4 und AP5. Diese Abbildung stammt aus einer Veröffentlichung von Kühn et al. [25].

Die Genotypisierung der 37204 SNPs wurde mittels des Illumina BovineSNP50 Beadchips durchgeführt und Genexpressionsprofile von 10069 Transkripten von Gewebeproben der

Nebennierenrinde wurden mittels des Affymetrix GeneChipBovine v1 Arrays gemessen. Die Daten wurden von Teilprojekt TP3.2 zur Verfügung gestellt. Zu Projektbeginn wurde in Zusammenarbeit mit diesem Teilprojekt ein Analyseplan aufgestellt. Auf diese Weise konnten Erfahrungen von früheren ähnlichen Analysen ausgetauscht und gebündelt werden. Dies war insbesondere deshalb notwendig, da aufgrund der familiären Struktur der untersuchten Rinderpopulation neben der Qualitätskontrolle und Vorverarbeitung der Daten zusätzliche Aufbereitung der Expressionsdaten notwendig war.

Denn eine wesentliche Voraussetzung des zuvor entwickelten adaptiven Tests ist die statistische Unabhängigkeit der Daten in den zu vergleichenden Gruppen. Dies ist bei Expressionsdaten von Rindern – anders als bei Menschen - üblicherweise nicht gegeben, da die Tiere alle aus kleinen Familien mit wenigen Vätern kommen. Deshalb ist es von zentraler Bedeutung, die Daten vor einer Assoziationsanalyse zu dekorrelieren.

Die Genexpressionsdaten wurden deshalb mit Hilfe einer Serie von univariaten Analysen (eine je Merkmal) unter Anwendung eines „Sire-Dam-Modells“ auf die systematischen Einflussgrößen Jahr-Saison (fixer Effekt), Schlachtalter in Tagen (lineare Regression) und die (additiv-genetische) Verwandtschaft (zufälliger Effekt) korrigiert. Dieses Modell wurde in Zusammenarbeit mit den Teilprojekten 3.2 und 1.4 entwickelt. Ein Sire-Dam-Modell zeichnet sich dadurch aus, dass in der additiv-genetischen Verwandtschaftsmatrix $\mathbf{A}$ nur die Eltern der F2-Tiere (P0- und F1-Individuen) berücksichtigt werden, welche selbst keine Beobachtungswerte besitzen. Phänotypisierungen liegen nur für die F2-Individuen vor. Die Zuordnung der Merkmalswerte der F2-Individuen zu den Eltern erfolgt über deren Verwandtschaft.

Das allgemeine lineare gemischte Modell für die Expressionsmessung eines Gens hat folgende Gestalt

$$y_i = s_j + \mathrm{b}\mathrm{x}_i + \frac{1}{2}(a_{iv} + a_{im}) + e_i,$$

y_i Beobachtung des i-ten Tieres ($\mathrm{i} = 1, \dots, 145$)

s_j Jahr-Saison-Effekt ($j = 1, \dots, 25$)

x_i Schlachtalter in Tagen des i-ten Tieres

b lineare Regression auf das Schlachtalter

a_{iv} additiv-genetischer Effekt des Vaters von Tier i (halber Zuchtwert)

a_{im} additiv-genetischer Effekt der Mutter von Tier i (halber Zuchtwert)

e_i (zufällige) Residuen

Der Vektor der zufälligen additiv-genetischen Effekte wird als normalverteilt mit $\mathrm{N}(\mathbf{0}, \frac{\mathbf{1}}{\mathbf{2}}\mathbf{A}\sigma_{\mathbf{a}}^{\mathbf{2}})$ angenommen, wobei $\mathbf{A}$ die additiv-genetische Verwandtschaftsmatrix (2*Kinshipmatrix) und σ_{a}^2 die additiv-genetische Varianz darstellt. Für die (zufälligen) Residuen e_i wird $\mathrm{N}(0, \sigma_{\mathrm{e}}^2)$ angenommen, wobei σ_{e}^2 die Restvarianz des Sire-Dam-Modells bezeichnet (enthält $\frac{\mathbf{1}}{\mathbf{2}}\sigma_{\mathbf{a}}^{\mathbf{2}}$). Die Residuen aus diesem Modell enthalten somit die Mendelian-Sampling-Effekte für jedes Tier als unkorrelierte genetische Komponente sowie eine Zufallsabweichung. Da die Residuen aus diesem Modell somit als statistisch

unabhängig betrachtet werden können wurden diese anstelle der rohen Expressionsdaten für die Assoziationsanalyse verwendet.

Um einen effizienten Umgang mit den Daten zu gewährleisten wurde die Software plink verwendet [26]. Der in R implementierte adaptive Test wurde dann jeweils für ein SNP/Transkript-Paar aufgerufen. Da die Analyse genomweit durchgeführt werden sollte („trans“), mussten die Prozeduren so implementiert werden, dass die Berechnungen der knapp 400 Millionen Assoziationstests auf einem Rechencluster mit 30 Prozessoren parallel durchgeführt werden konnten, da diese auf einem PC mit einem Prozessor etwa 50 Tage gedauert hätten. Die zur Berechnung genutzten Server wurden aus Projektmitteln finanziert.

Um die große Menge an Ergebnissen effizient verwalten zu können erwies sich eine Datenbank als hilfreich. Die genomweite SNP-Expressions-Assoziationsanalyse wurde dann wie geplant durchgeführt. Die im Rahmen dieser Analyse berechneten p-Werte können transformiert und als Gewichte für genom-weite SNP/Transkriptom-Analysen aus „OMICS“-Netzwerkstrukturen verwendet werden.

Die in AP4 gesteckten Ziele wurden also wie geplant erreicht. Da durch die geschickte und flexible Implementierung der Verfahren nun eine komplette genomweite Assoziationsanalyse molekularer Hochdurchsatzdaten in wenigen Tagen möglich ist konnten diese Methoden sogar ohne großen Mehraufwand dazu verwendet werden, Assoziationen zwischen SNPs und Metabolom-Messungen derselben Rinderpopulation zu identifizieren.

AP5

In diesem Arbeitspaket stand die Aufklärung der molekularen Funktion gefundener genetischer Biosignaturen im Vordergrund. Das zuvor entwickelte adaptive Verfahren ermöglicht die Identifizierung von genetischen Varianten, die einen Einfluss auf die Expression bestimmter Gene haben und somit auf den jeweiligen Phänotyp wirken.

Zunächst mussten sowohl die SNPs als auch die Transkripte annotiert werden. Auch hier fungierte eine Datenbank als hilfreiches Werkzeug. Um SNP-Transkript-Assoziationen identifizieren zu können mussten die in AP4 berechneten p-Werte für multiples Testen adjustiert werden. Abhängig von der relativen Position von SNP und Gen im Genom wurden unterschiedliche Signifikanz-Grenzen auf Basis der Bonferroni-Prozedur festgelegt. Für Paare die in cis liegen – d.h. SNP und exprimiertes Gen liegen beide innerhalb eines Fensters von einer Million Basenpaaren – wurde ein Schwellenwert von $5*10^{-5}$ verwendet und für Paare in trans – wo also der SNP nicht mehr innerhalb des Fensters, vielleicht sogar auf einem anderen Chromosom positioniert ist als das assoziierte Gen – ein Schwellenwert von $5*10^{-8}$. Insgesamt konnten 613 SNPs identifiziert werden, welche die Expression von 249 Genen der Nebennierenrinde des Rindes signifikant beeinflussen. Diese Assoziationen sind in Abbildung 8 dargestellt. In dieser Abbildung ist leicht erkennbar, dass die meisten identifizierten Assoziation nahe der Hauptdiagonalen – also in cis - liegen und nur wenige in trans.

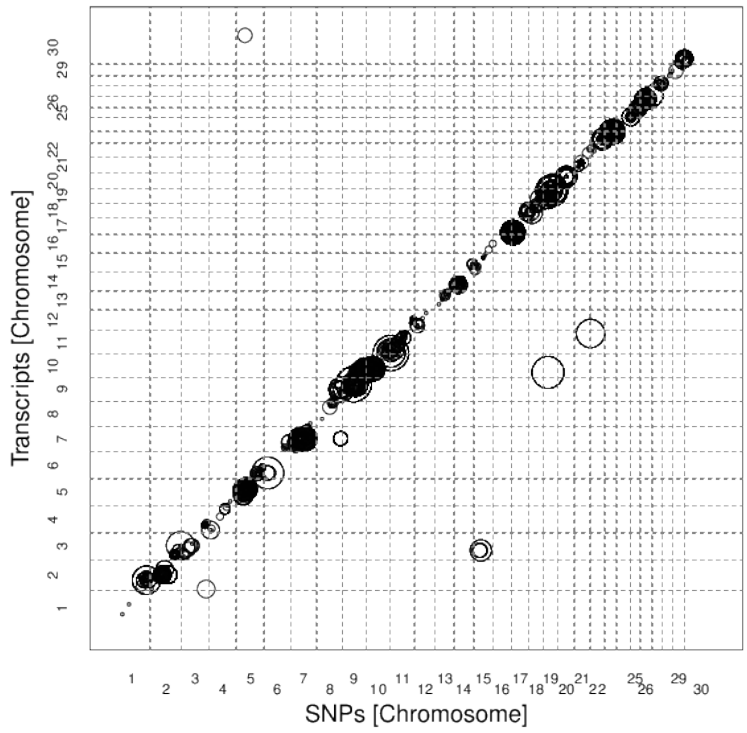

Abbildung 8: Zentrale Resultate der eQTL-Analyse. Signifikante SNP-Expressions-Assoziationen sind als Kreise bezüglich ihrer Position im Genom dargestellt. Die Größe der Kreise steht in reziproker Relation zum entsprechenden p-Wert.

Bei der Auswertung der Ergebnisse fand wiederum eine enge Zusammenarbeit mit Teilprojekt TP3.2 statt. Ein Vergleich mit der online-Datenbank QTLDB (http://www.animalgenome.org/cgi-bin/QTLdb/BT/index) zeigte mehrere Überschneidungen der hier identifizierten eQTL der Nebennierenrinde mit zuvor berichteten QTL anderer Produktionsmerkmale aus anderen Rinderpopulationen. Außerdem liegen laut Aussagen von TP3.2 sowohl ein Teil der identifizierten SNPs als auch der Transkripte in chromosomalen Regionen, die mit bestimmte Verhaltensmerkmalen assoziiert sind, welche an derselben Kreuzpopulation untersucht wurden.

Die assoziierten Gene sollten auch mit Hilfe von Pathway-Analyse-Software näher untersucht werden, um auch eine biologische Relevanz in ausgewählten biologischen Netzwerken aufzuklären. Dies wurde mit Ingenuity® Pathway Analysis (IPA® , QIAGEN Redwood City, www.qiagen.com/ingenuity) durchgeführt.

Molecular and Cellular Functions

Name	p-value	# Molecules
Carbohydrate Metabolism	3,66E-04 - 4,38E-02	11
Nucleic Acid Metabolism	3,66E-04 - 3,60E-02	6
Small Molecule Biochemistry	3,66E-04 - 4,38E-02	15
Lipid Metabolism	7,27E-04 - 4,38E-02	9
Cell Death and Survival	2,25E-03 - 4,64E-02	9

Tabelle 1: Die wichtigsten regulatorischen Netzwerke der Ingenuity® Pathway Analysis.

Die wichtigsten in dieser Analyse identifizierten funktionalen Netzwerke sind Tabelle 1 zu entnehmen. Drei dieser Netzwerke, welche teilweise in Abbildung 9 dargestellt sind, hängen mit dem Stoffwechsel von Kohlenhydraten, Nukleinsäuren und Fetten zusammen. Dieses Ergebnis ist sehr plausibel. Denn laut der mendelschen Spaltungsregel sind bei einer Kreuzung von homogenen Individuen in einem dominant-rezessive Vererbung durchschnittlich ein Viertel der F2-Individuen reinerbig mit zwei rezessiven Erbanlagen, ein weiteres Viertel reinerbig mit dominanten Erbanlagen und die Hälfte mit heterozygoten Erbanlagen. Insgesamt besteht also im Phänotyp ein Verhältnis von

3:1, im Genotyp eines von 1:2:1. Bei den hier analysierten Daten handelt es sich ja um Messungen der Expression bei einer F2 Kreuzpopulation. Bei den Vorfahren aus der F0 Generation, welche aus Fleischrindern und Milchkühen bestehen, ist von unterschiedlichen homozygoten Genotypen im Nebennierengewebe auszugehen, welches einen großen Einfluss auf den Stoffwechsel hat. Da sich infolgedessen auch die Expressionsniveaus der betreffenden Gene in der F2 gemäß obiger Regel aufteilen müsste, war zu erwarten, dass ein großer Teil der identifizierten eQTLs mit dem Stoffwechsel in Zusammenhang stehen. Dass dies tatsächlich der Fall ist, spricht für die Validität der identifizierten Assoziationen und somit auch für die Validität der verwendeten Analyse-Methoden insgesamt.

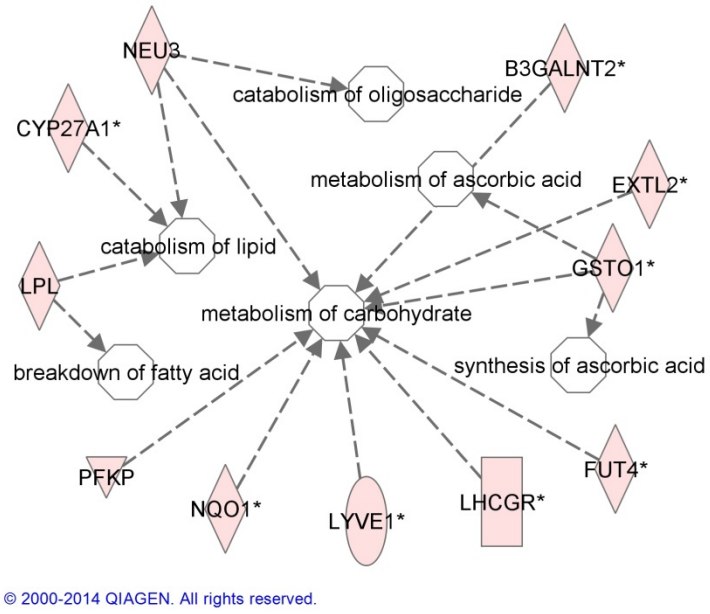

Abbildung 9: Auszug der identifizierten regulatorischen Netzwerke bei der F2 Kreuzpopulation, die mit dem Stoffwechsel beeinflussen.

Eingehendere Analysen von Produktions- und Verhaltensmerkmalen in der untersuchten Kreuzpopulation z.B. anhand von genomweiten Assoziationsstudien (GWAs) wird sicherlich in Zukunft noch mehr Einsicht in die molekularen Mechanismen bringen, durch welche die in diesen Analysen identifizierten eQTL die entsprechenden Merkmale beeinflussen. Die in den Arbeitspaketen AP4 und AP5 erzielten Ergebnisse wurden als Poster auf der internationalen Tagung IGES 2014 vorgestellt. Darüber hinaus ist eine Veröffentlichung in einer internationalen Fachzeitschrift in Zusammenarbeit mit Teilprojekt TP3.2 geplant.

2. Wichtigsten Positionen des zahlenmäßigen Nachweises

Personalkosten

Alle in diesem Projekt durchgeführten Arbeiten konnten nur von einem erfahrenen Biometriker durchgeführt werden, der Erfahrung mit der Durchführung von Monte-Carlo-Simulationsstudien sowie der Analyse von SNP-Expressions-Assoziationsstudien hat. Deshalb wurde wie beantragt ein geeigneter wissenschaftlicher Mitarbeiter aus Projektmitteln finanziert.

Anschaffungen

Da die Dell PowerEdge M600 Blade Server, die zur Verarbeitung der anfallenden Datenmengen benötigt wurden, zum Zeitpunkt der Bestellung nicht mehr hergestellt wurden, wurden anstelle der drei oben genannten Einheiten zwei Power Edge M620 Blade Server beschafft, deren Kosten die beantragten Mittel nicht überstiegen. Darüber hinaus wurden wie beantragt Festplatten und Speicherbänder zur Datensicherung angeschafft.

Reisekosten

Aus den zur Verfügung gestellten Mitteln wurden Reisen zu den Biometrischen Kolloquien 2013 und 2014 (Freiburg und Bremen) finanziert, um die gewonnenen Erkenntnisse vorzustellen. Außerdem wurden Reisen zum Kooperationspartner nach Dummerstorf getätigt.

3. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Wie unter den Punkten 1. und 2. aufgeschlüsselt, wurden die Arbeiten und die finanziellen Mittel entsprechend der Arbeitspakete ausgerichtet. Generell scheinen adaptive statistische Tests im Bereich der Identifikation von Biomarkern bislang kaum genutzt zu werden. Deshalb sind die Ergebnisse aus diesem Teilprojekt potentiell sehr wertvoll für zukünftige Forschungsprojekte, die durch diese Arbeiten auf die Anwendung flexibler statistischer Verfahren aufmerksam werden.

Auszug aus einem der Gutachterkommentare des Journals Methods of Information in Medicine: „The raised question is truly important and the paper addresses a current very relevant topic because gene, protein, or metabolite expression levels are often non-normally distributed, standard statistical approaches may fail in this situation and permutation testing is not an efficient option in this situation. The paper represents a sound methodical contribution to a special topic and is of fundamental practical relevance. Altogether, the manuscript is very carefully written and scientifically of high rank.“

Dieser und andere Kommentare von dritten, z.B. auf Fachkonferenzen als Reaktion auf die Präsentation der erzielten Ergebnisse von AP2 und AP3, machen deutlich, dass die Benutzung von adaptiven Testverfahren für Assoziationsanalysen von Genotypen mit quantitativen Phänotypen wie zum Beispiel Expressionsdaten ein vielversprechender Ansatz ist, und dass auch die Ergebnisse, die schon durch dieses Teilprojekt erzielt werden konnten von nicht unerheblicher wissenschaftlicher sowie praktischer Relevanz sind. Die biologisch plausiblen Ergebnisse der eQTL-Analyse in den Arbeitspaketen AP4 und AP5 sprechen ebenfalls für die Anwendbarkeit des hier entwickelten adaptiven Tests in vergleichbaren Studien zur Identifizierung von Biomarkern z.B. in SNP-Expressions-Assoziationsstudien.

4. Verwertbarkeit der Ergebnisse

Die Verwertbarkeit der Ergebnisse ist in vollem Umfang gegeben. Die Verwertung der wissenschaftlichen Projektergebnisse erfolgte wie vorgesehen primär durch Publikationen in international anerkannten Zeitschriften mit Begutachtungssystem. Des Weiteren wurden die Ergebnisse durch Vorträge und Poster auf Fachtagungen und Projekttreffen der Fachöffentlichkeit dargestellt. Das entwickelte Software-Paket mit dem adaptiven Test wurde schon frühzeitig mehreren Kooperationspartnern zur Verfügung gestellt und kann nach wie vor auf Anfrage hin kostenfrei weitergegeben werden.

5. Fortschritt bei anderen Stellen

Keine bekannt.

6. Erfolgte oder geplante Veröffentlichungen

Tagungsbeiträge

- European Mathematical Genetics Meeting (EMGM 2012, Göttingen, 12.-13.04.2012)
  Scheinhardt, M.O.; Szymczak, S.; Ziegler, A.: Adaptive linear rank tests for eQTL studies (Vortrag)

- Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS 2012, Braunschweig, 16.-21.09.2012)
  Scheinhardt, M.O.; Szymczak, S.; Ziegler, A.: Adaptive Tests für Biomarker-Studien (Vortrag)
- Annual Scientific International Genetic Epidemiology Society Meeting (IGES 2014, Wien, 28.-30.08.2014)
  Scheinhardt, M.O.; Brand, B.; Zimmer, D.; Reinsch, N.; Schwerin, M.; Ziegler, A.: EQTL and pathway analysis on expression profiles of a cattle cross (Poster)

Publikationen in wissenschaftlichen Zeitschriften

- Szymczak, S., et al., *Adaptive linear rank tests for eQTL studies.* Statistics in Medicine, 2013. **32**(3): p. 524-537.
- Scheinhardt, M.O. and A. Ziegler, *Location Tests for Biomarker Studies: A Comparison Using Simulations for the Two-sample Case.* Methods of Information in Medicine, 2013. **52**(4): p. 351-359.

Eine weitere Publikation über die Ergebnisse der Arbeitspakete AP4 und AP5 in Zusammenarbeit mit Teilprojekt TP3.2 ist geplant.

III. Erfolgskontrollbericht

Siehe Anlage

IV. Berichtsblatt

1. ISBN oder ISSN	2. Berichtsart **Schlussbericht**
3. Titel des Berichts AgroCluster: PHÄNOMICS – Ein systembiologischer Ansatz zur Genotyp-Phänotyp-Abbildung im Kontext von Leistung, Gesundheit und Wohlbefinden bei den Nutztieren Rind und Schwein – (Teilprojekt TP 1.3)	
4. Autor(en) des Berichts [Name(n), Vorname(n)] Ziegler, Andreas; Scheinhardt, Markus Oliver	5. Abschlussdatum des Vorhabens Juni 2014 6. Veröffentlichungsdatum Dezember 2014 7. Form der Publikation Schlussbericht
8. Durchführende Institution(en) (Name, Adresse) Universität zu Lübeck Medizinische Fakultät Institut für Medizinische Biometrie und Statistik Ratzeburger Allee 160, Haus 24 23562 Lübeck	9. Ber. Nr. Durchführende Institution 10. Förderkennzeichen 0315536F 11. Seitenzahl Bericht 20
12. Fördernde Institution (Name, Adresse) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. Literaturangaben 26 14. Tabellen 1 15. Abbildungen 9
16. Zusätzliche Angaben	
17. Vorgelegt bei (Titel, Ort, Datum)	
18. Kurzfassung SNP-Genexpressions-Assoziationen sind – sowohl beim Menschen als auch bei Nutztieren - von zunehmend großer Bedeutung, um molekulare Funktionen genetischer Variation die mit interessanten Phänotypen assoziiert sind, aufzuklären. Standardverfahren führen jedoch in vielen Situationen zu liberalen Ergebnissen, insbesondere bei schiefen Verteilungen oder Verteilungen mit Ausreißern. Ziel des Teilprojekts ist es daher, einen adaptiven statistischen Test zu entwickeln, der auch für nichtnormalverteilte Daten statistisch robust ist und trotzdem eine möglichst große statistische Macht hat. Zunächst wurde mittels einer systematischen Literaturrecherche ein geeignetes etabliertes Verfahren ausgewählt, welches dann unter Verwendung von Monte-Carlo-Simulationsstudien für den Einsatz in SNP-Genexpressions-Assoziationen modifiziert und optimiert wurde. Anschließend wurde anhand von Simulationsstudien ein Vergleich der Eigenschaften des neuen Tests mit anderen adaptiven Tests und Standardtests durchgeführt. Schließlich wurde das erfolgreich evaluierte Verfahren im Rahmen des Kompetenznetzes PHÄNOMICS als Teil des Verbundprojektes VP1 - Integrative Bioinformatik - auf Daten aus einer F2-Kreuzpopulation von Deutschen Holstein-Kühen (Milchkuh) und Charolais-Bullen (Fleischrind) eingesetzt. Viele der auf diese Weise identifizierten Assoziationen hängen mit dem Stoffwechsel von Kohlenhydraten, Nukleinsäuren und Fetten zusammen, was laut der mendelschen Spaltungsregel biologisch plausibel ist. Sowohl der Vergleich der Verfahren anhand von Simulationsstudien als auch die erfolgreiche Anwendung auf realen Daten sprechen für den Einsatz des modifizierten adaptiven Tests in ähnlichen Studien zur Identifizierung von Biomarkern bei potentiell nichtnormalverteilten Daten.	
19. Schlagwörter SNP-Expressions-Assoziationsstudie, eQTL-Studie, adaptiver Test, Simulationsstudie	
20. Verlag	21. Preis

Literaturverzeichnis

1. Szymczak S, Igl BW, Ziegler A. Detecting SNP-expression associations: A comparison of mutual information and median test with standard statistical approaches. Statistics in Medicine 2009; 28: 3581-3596.
2. Büning H. Adaptive Tests for the c-sample Location Problem -- The Case of Two-sided Alternatives. Communications in Statistics -- Theory and Methods 1996; 25: 1569-1582.
3. Beier F. Adaptive Tests bei nicht-monotonen Dosis-Wirkungsbeziehungen: Universität Dortmund; 1996.
4. Hogg RV, Fisher DM, Randles RH. A Two-sample Adaptive Distribution-free Test. Journal of the American Statistical Association 1975; 70: 656-661.
5. O'Gorman TW. An adaptive test for the one-way layout. The Canadian Journal of Statistics 1997; 25: 269-279.
6. Keselman H, Wilcox RR, Lix LM, Algina J, Fradette K. Adaptive robust estimation and testing. British Journal of Mathematical and Statistical Psychology 2007; 60: 267-293.
7. Welch B. On the comparison of several mean values: An alternative approach. Biometrika 1951; 38: 330-336.
8. Hogg RV. Adaptive robust procedures: A partial review and some suggestions for future applications and theory. Journal of the American Statistical Association 1974; 69: 909-923.
9. Hogg RV. A new dimension to nonparametric tests. Communications in Statistics - Theory and Methods 1976; 5: 1313 - 1325.
10. Hogg RV, Lenth RV. A review of some adaptive statistical techniques. Communications in Statistics -- Theory and Methods 1984; 13: 1551-1579.
11. Hušková M. Adaptive methods. Handbook of Statistics 1984; 4: 347-358.
12. Büning H. Robustness and Power of Parametric, Nonparametric, Robustified and Adaptive Tests -- The Multi-sample Location Problem. Statistical Papers 2000; 41: 381-407.
13. Hüsler J. On the Two-sample Adaptive Distribution-free Test. Communications in Statistics -- Simulation and Computation 1987; 16: 55-68.
14. Sun S. A class of adaptive distribution-free procedures. Journal of Statistical Planning and Inference 1997; 59: 191-211.
15. Neuhauser M, Büning H, Hothorn L. Maximum test versus adaptive tests for the two-sample location problem. Journal of applied statistics 2004; 31: 215-227.
16. Kössler W, Kumar N. An adaptive test for the two-sample location problem based on U-statistics. Communications in Statistics -- Simulation and Computation 2008; 37: 1329-1346.
17. Kössler W, Lesener WF. Adaptive Lokationstests mit U-Statistiken. 2010.

18. Ruberg SJ. A continuously adaptive nonparametric two-sample test. Communications in Statistics - Theory and Methods 1986; 15: 2899-2920.
19. O'Gorman TW. An adaptive two-sample test based on modified Wilcoxon scores. Communications in Statistics -- Simulation and Computation 1996; 25: 459-479.
20. Wells CS, Hintze JM. Dealing with assumptions underlying statistical tests. Psychology in the Schools 2007; 44: 495-502.
21. Rayner G, MacGillivray H. Weighted quantile-based estimation for a class of transformation distributions. Computational Statistics and Data Analysis 2002; 39: 401-433.
22. Bradley JV. Robustness? British Journal of Mathematical and Statistical Psychology 1978; 31: 144-152.
23. Szymczak S, Scheinhardt MO, Zeller T, Wild PS, Blankenberg S, Ziegler A. Adaptive linear rank tests for eQTL studies. Statistics in Medicine 2013; 32: 524-537.
24. Scheinhardt M, Ziegler A. Location Tests for Biomarker Studies: A Comparison Using Simulations for the Two-sample Case. Methods of Information in Medicine 2013; 52: 351-359.
25. Kühn C, Bellmann O, Voigt J, Wegner J, Guiard V, Ender K. An experimental approach for studying the genetic and physiological background of nutrient transformation in cattle with respect to nutrient secretion and accretion type. Arch Anim Breed 2002; 45: 317-330.
26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 2007; 81: 559-575.