

Aus dem Institut für Medizinische Biometrie und Statistik
der Universität zu Lübeck
Direktor: Univ.-Prof. Dr. rer. nat. Andreas Ziegler

Adaptive und Mutual-Information-basierte Verfahren für eQTL-Studien

Inauguraldissertation
zur
Erlangung der Doktorwürde
der Universität zu Lübeck

Aus den Sektionen Informatik/Technik und Naturwissenschaften

vorgelegt von
Silke Szymczak
aus Bonn

Lübeck 2011

1. Berichterstatter: Univ.-Prof. Dr. rer. nat. Andreas Ziegler
2. Berichterstatter: Univ.-Prof. Dr. rer. nat. Heinz Handels
Tag der mündlichen Prüfung: 2. Februar 2011

Zum Druck genehmigt.

Lübeck, den 7. April 2011

Inhaltsverzeichnis

Abbildungsverzeichnis	iv
Tabellenverzeichnis	vi
1. Einleitung	1
1.1. Motivation	1
1.2. eQTL-Studien	3
1.3. Statistisches Testszenario	6
1.4. Zielsetzung	7
1.5. Aufbau der Arbeit	8
2. Simulationsstudie I: Klassische und Mutual-Information-basierte Verfahren	10
2.1. Zielsetzung	10
2.2. Methoden	11
2.2.1. Klassische Verfahren	11
2.2.2. Mutual-Information-basierte Verfahren	15
2.3. Simulationen	19
2.3.1. Simulationsszenario	19
2.3.2. Simulierte Verteilungen	21
2.3.3. Parameter	23
2.4. Ergebnisse	23
2.4.1. Empirische Fehler 1. Art	23
2.4.2. Empirische Güte	24
2.5. Diskussion	25
3. Simulationsstudie II: Adaptive Verfahren	29
3.1. Zielsetzung	29

3.2. Methoden	30
3.2.1. Selektorstatistiken	31
3.2.2. Lineare Rangstatistiken	34
3.2.3. Adaptive lineare Rangtests	38
3.3. Simulationen	40
3.3.1. Verteilungsfamilien	41
3.3.2. Simulierte Verteilungen	43
3.3.3. Parameter	44
3.4. Ergebnisse	45
3.4.1. Entwicklung eines neuen adaptiven linearen Rangtests . . .	46
3.4.2. Vergleich der adaptiven Tests	51
3.5. Diskussion	57
4. Realdaten: Gutenberg-Herz-Studie	58
4.1. Zielsetzung	58
4.2. Daten	59
4.3. Methoden	59
4.3.1. Vorverarbeitung der Genotypdaten	60
4.3.2. Vorverarbeitung der Expressionsdaten	61
4.3.3. Assoziationen zwischen SNPs und Transkripten	62
4.4. Ergebnisse	63
4.4.1. Vergleich der beiden Kohorten	63
4.4.2. Vergleich der signifikanten und replizierbaren Assoziationen	64
4.4.3. Zusammenhang mit Ergebnissen aus GWA-Studien	66
4.5. Diskussion	67
5. Diskussion und Ausblick	71
6. Zusammenfassung	74
A. Anhang	76
A.1. Entscheidungsgrenzen für adaptive lineare Rangtests	76
A.2. Empirische Fehler 1. Art für gleiche Verteilungsformen	77
A.2.1. Lineare Rangtests	77
A.2.2. Adaptive Tests	79
A.3. Optimale lineare Rangtests mit neuem Klassifikationsschema . . .	81

Literaturverzeichnis	82
Danksagung	96

Abbildungsverzeichnis

1.1. Ablauf einer eQTL-Studie	4
2.1. Simulationsszenarien für schiefe Verteilungen	22
3.1. Gewichte für verschiedene lineare Rangtests	36
3.2. Adaptive lineare Rangtests von Büning und Beier	39
3.3. Dichten der g -und- k -Verteilungen	42
3.4. Theoretische Selektorstatistiken für die simulierten gleichen Verteilungsformen	44
3.5. Simulationsszenarien für schiefe Verteilungen	45
3.6. Empirische Güte der linearen Rangtests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,5	47
3.7. Empirische Güte der linearen Rangtests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,3	48
3.8. Optimale lineare Rangtests für Verteilungen mit gleicher Verteilungsform	49
3.9. Neues adaptives Verfahren	50
3.10. Kombination des Median-Tests und des neuen adaptiven Tests . . .	51
3.11. Empirische Güte der adaptiven Tests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,5	52
3.12. Empirische Güte der adaptiven Tests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,3	53
4.1. Venn-Diagramm	65
4.2. Regionale Assoziationsabbildung	70
A.1. Empirische Fehler 1. Art der linearen Rangtests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,5	77

A.2. Empirische Fehler 1. Art der linearen Rangtests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,3	78
A.3. Empirische Fehler 1. Art der adaptiven Tests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,5	79
A.4. Empirische Fehler 1. Art der adaptiven Tests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,3	80
A.5. Optimale lineare Rangtests für Verteilungen mit gleicher Verteilungsform	81

Tabellenverzeichnis

2.1. Empirische Fehler 1. Art für Mutual-Information-basierte und klassische Verfahren	27
2.2. Empirische Güte für Mutual-Information-basierte und klassische Verfahren	28
3.1. Empirische Fehler 1. Art für adaptive Verfahren	55
3.2. Empirische Güte für adaptive Verfahren	56
4.1. Vergleich der beiden Kohorten	63
4.2. Anzahl signifikanter und replizierter Assoziationen und Transkripte	64
4.3. Signifikante Assoziationen aus GWA-Studien	69
A.1. Adaptiver linearer Rangtest von Büning	76
A.2. Adaptiver linearer Rangtest von Beier	76

1. Einleitung

1.1. Motivation

Das Erkrankungsrisiko für komplexe Erkrankungen, wie Herzinfarkt, Diabetes oder Krebs, wird neben Umweltfaktoren auch durch genetische Varianten beeinflusst. Um solche genetischen Komponenten zu identifizieren, wurden in den letzten Jahren viele genomweite Assoziationsstudien (GWA-Studien) durchgeführt. Hierbei werden jedoch nicht wie bei Kandidatengenstudien Varianten in biologisch interessanten Genen oder Stoffwechselwegen untersucht. Stattdessen wird möglichst die gesamte Variabilität im Genom erfasst. Mit den heutigen Technologien ist es möglich, zwischen 500.000 und bis zu einer Million Einzel-Nukleotid-Polymorphismen (SNPs, engl. single nucleotide polymorphisms) auf einem einzelnen sogenannten SNP-Array zu genotypisieren. Jeder dieser SNPs wird einzeln auf Assoziation mit dem klinischen Phänotyp wie beispielsweise Herzinfarkt getestet, d.h. es werden SNPs gesucht, bei denen eine bestimmte Variante bei Erkrankten häufiger vorkommt als bei Gesunden.

Mittlerweile wurden mit solchen Assoziationsstudien mehr als 150 SNPs für über 60 verschiedene komplexe Erkrankungen und andere biologische Phänotypen identifiziert [64]. Nur in wenigen Fällen ist etwas über den molekularen Mechanismus bekannt, der einen Hinweis auf die Entstehung der Krankheit liefern kann. Der Grund hierfür ist, dass die meisten in GWA-Studien identifizierten SNPs nicht in proteinkodierenden Bereichen, sondern in der Nähe verschiedener Gene oder sogar in sogenannten „genomischen Wüsten“ liegen, d.h. in Regionen außerhalb bekannter kodierender Sequenzen oder regulatorischer Bereiche [38]. Ein möglicher molekularer Mechanismus für diese SNPs ist die Beeinflussung

der Expression benachbarter oder auch weiter entfernt liegender Gene. Beispielsweise sind mehrere SNPs in einer „genomischen Wüste“ auf Chromosom 5p13.1 mit Morbus Crohn assoziiert. Diese SNPs haben einen Einfluss auf das Expressionsniveau von *PTGER4*, obwohl dieses Gen ungefähr 270 kb von den Assoziationssignalen entfernt ist [60]. In einer anderen GWA-Studie zu Asthma konnte gezeigt werden, dass mit der Krankheit assoziierte SNPs auf Chromosom 17q23 nur mit der Expression eines einzelnen Gens (*ORMDL3*) von mehreren Genen in dieser Region assoziiert sind [69].

Dieser Einfluss auf die Expression ist möglich durch veränderte regulatorische Sequenzen, die sich entweder in der Nähe des Gens oder an anderer Stelle im Genom befinden. Neben diesen Sequenzen, die die Transkription beeinflussen, gibt es regulatorische RNAs (engl. ribonucleic acid), die beispielsweise durch RNA-Interferenz sequenzspezifisch mRNAs (engl. messenger RNA) abbauen und somit die Translation verhindern. Mittlerweile wurde beispielsweise in der Chromosom 9p21.3 Region eine nichtkodierende RNA (*ANRIL*) identifiziert [73], die in für Arteriosklerose wichtigem Gewebe exprimiert wird und deren Expression mit SNPs in dieser Region assoziiert ist [45]. Über RNA-Interferenz-Mechanismen ist jedoch nichts bekannt.

In sogenannten eQTL-Studien (engl. expression quantitative trait loci), auch SNP-Expressions-Assoziationsstudien oder „genetical genomics“ genannt, werden Assoziationen zwischen SNPs und Expressionswerten von mehreren zehntausend Transkripten untersucht. Hierbei werden also SNPs identifiziert, bei denen das typische Expressionsniveau eines bestimmten Gens vom Genotyp abhängt. Zur Bestimmung der Expression möglichst aller bekannten proteinkodierenden Gene (Transkriptom-Analyse) werden Microarrays verwendet, die über Hybridisierung und Markierung mit Fluoreszenzfarbstoffen die Stärke der Expression bestimmen.

Die vorliegende Arbeit befasst sich mit Mutual-Information-basierten und adaptiven statistischen Methoden zur Identifizierung solcher Assoziationen zwischen SNPs und Transkripten. Im Abschnitt 1.2 wird zunächst den Ablauf einer eQTL-Studie beschrieben, bevor im Abschnitt 1.3 näher auf das statistische Testscenario

mit der in dieser Arbeit betrachteten Nullhypothese eingegangen wird. Anschließend werden die Zielsetzung und der Aufbau dieser Arbeit näher erläutert.

1.2. eQTL-Studien

Die zellspezifische Genexpression wird sowohl durch Umwelteffekte als auch durch genetische Komponenten beeinflusst [17]. Letztere können in eQTL-Studien identifiziert werden, deren typischer Ablauf in Abbildung 1.1 dargestellt ist.

Abhängig von der Fragestellung werden zunächst die Studienteilnehmer ausgewählt. In eQTL-Studien, die allgemein die genetische Regulation der Genexpression untersuchen, werden meist gesunde Probanden verwendet [23, 87], wie beispielsweise Personen aus dem HapMap-Projekt [66], für die mittlerweile Genotypdaten von mehr als 3 Millionen SNPs öffentlich verfügbar sind. Andere Studien, die an krankheitsspezifischen Assoziationen interessiert sind, untersuchen hingegen beispielsweise Fälle mit Zöliakie [35, 44] oder Alzheimer [97].

In diesen Probanden werden mittels Array-Technologie sowohl Genotypen von genomweit verteilten SNPs als auch die Expression möglichst vieler bekannter Transkripte, d.h. mRNAs, bestimmt. Für die Bestimmung der Genotypen ist der Zelltyp nicht relevant, da bis auf wenige Ausnahmen alle Zellen die gleiche DNA besitzen. Hingegen ist die Genexpression sehr stark vom jeweiligen Gewebe und Zelltyp abhängig, so dass in einer eQTL-Studie nur zelltypspezifische eQTLs identifiziert werden können. HapMap-basierte Studien verwenden Epstein-Barr-Virus transformierte lymphoblastoide Zelllinien, die jedoch einen bedeutenden Nachteil aufweisen. Durch die Immortalisierung und fortwährende Kultur der Zellen kann es zu Änderungen in der Genexpression kommen, so dass die gefundenen Assoziationen womöglich in vivo keinen Einfluss haben. Alternativ wird die Expression beispielsweise in verschiedenen Typen von Blutzellen [31, 35, 44] sowie Leber- [82] oder Gehirngewebe [97] analysiert.

Vor den statistischen Analysen ist eine strikte Qualitätskontrolle der Genotyp- so-

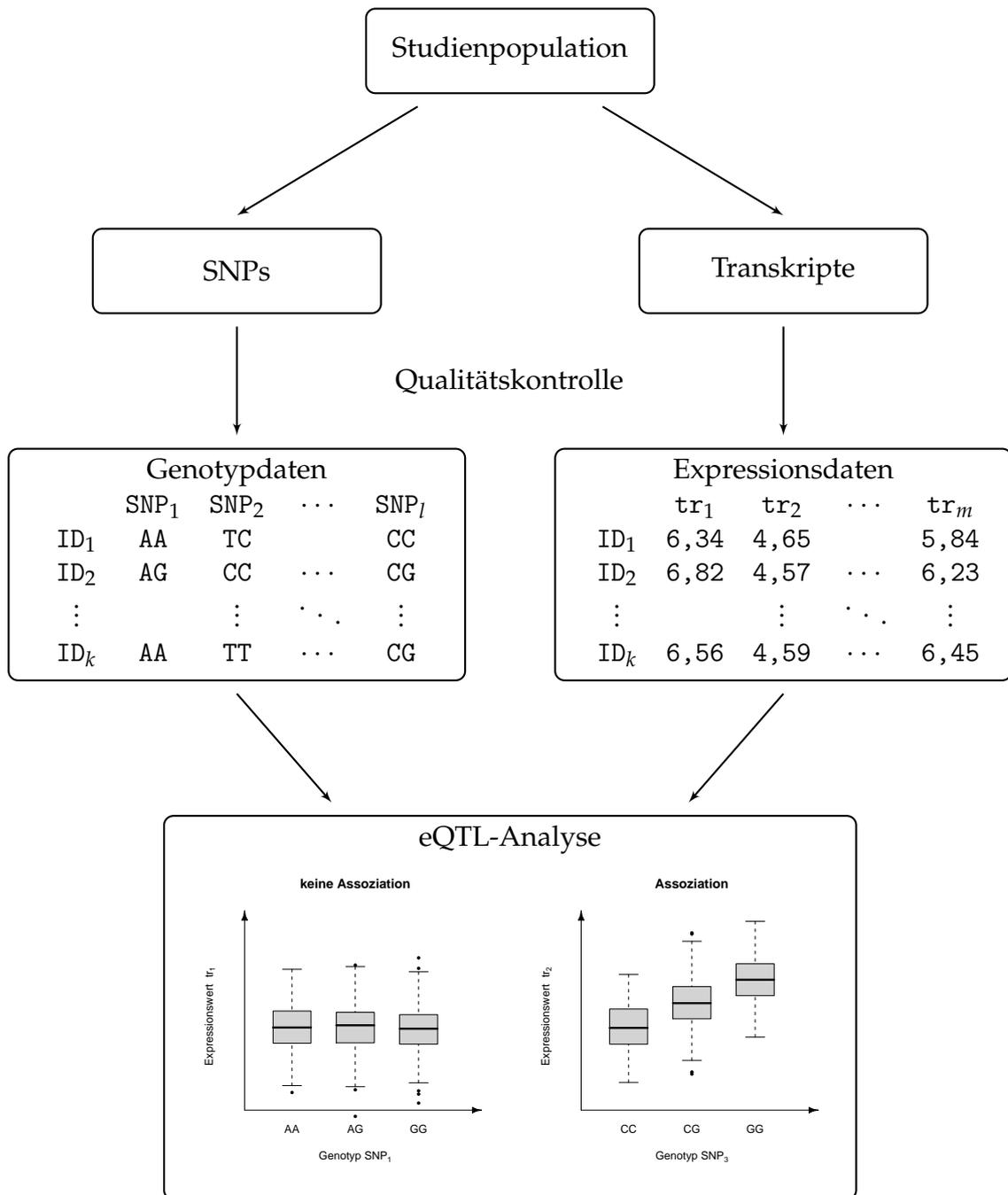


Abbildung 1.1: Ablauf einer eQTL-Studie. Für jeden Probanden der Studienpopulation werden sowohl Genotypen von genomweit verteilten SNPs als auch die Expressionsniveaus von möglichst vielen Transkripten in einem relevanten Zelltyp bestimmt. Nach intensiver Qualitätskontrolle liegen für k Probanden Genotypen von l SNPs sowie vorverarbeitete Expressionswerte für m Transkripte vor. Für jede Kombination der l SNPs und m Transkripte wird ein Test auf Assoziation durchgeführt (eQTL-Analyse). Im linken Boxplot sind die Expressionswerte für alle Genotypen sehr ähnlich; es liegt also keine Assoziation vor. Hingegen zeigt der rechte Boxplot ein Beispiel für eine Assoziation, bei der die Expression mit der Anzahl der G-Allele ansteigt.

wie der Expressionsdaten notwendig, um problematische SNPs oder Transkripte zu identifizieren (s. auch Abschnitt 4.3). Für SNPs haben sich mittlerweile eine Reihe von Kriterien etabliert [95, 105], anhand derer solche SNPs ausgeschlossen werden, die eine zu geringe Allelhäufigkeit oder einen hohen Anteil an fehlenden Genotypen haben bzw. deren Genotyphäufigkeiten vom Hardy-Weinberg-Gleichgewicht abweichen. Auf Expressionsebene werden in den meisten Fällen nur solche Transkripte für weitere Analysen verwendet, deren Expressionsniveaus größer als das Hintergrundrauschen sind. Aufgrund der hybridisierungs-basierten Array-Technologie kann es außerdem zu unechten Assoziationen zwischen SNPs innerhalb eines Gens und des jeweiligen Gens kommen, wenn SNPs in der für die Hybridisierung verwendeten Sonde zu unterschiedlichem Bindungsverhalten führen [1]. In manchen Studien werden solche Transkripte deshalb ausgeschlossen.

Nach der Qualitätskontrolle liegen für jeden Probanden diskrete Genotypdaten für alle qualitätskontrollierten SNPs sowie quantitative (logarithmierte) Expressionsdaten für alle detektierbaren Transkripte vor. Für jede mögliche Kombination von SNPs und Transkripten wird ein statistischer Test auf Assoziation durchgeführt. Hierbei werden für die drei Genotypen des jeweiligen SNPs die mittleren oder durchschnittlichen Expressionsniveaus des Transkripts miteinander verglichen. Wenn sich wie im rechten Boxplot in Abbildung 1.1 die mittleren Expressionswerte eines Gens bei Trägern verschiedener Genotypen stark voneinander unterscheiden, liegt eine Assoziation vor. Alternativ kann die Expression wie im linken Boxplot auch unabhängig vom Genotyp sein.

Abhängig von der genomischen Position der assoziierten SNPs zu dem jeweiligen Gen werden *cis*- und *trans*-Assoziationen unterschieden. Beim erstgenannten Typ liegen die SNPs innerhalb bzw. in einer begrenzten Region um das jeweilige Gen. Dieser Bereich beträgt in den meisten Studien 500 kb bis 1 Mb, gemessen vom Start und Ende des Gens. Solche SNPs in *cis* beeinflussen wahrscheinlich die Expression über regulatorische Elemente wie Transkriptionsfaktor-Bindungsstellen oder Enhancer-Elemente. Hingegen sind in *trans*-Assoziationen die SNPs weit vom Gen entfernt oder liegen sogar auf einem anderen Chromosom. Solche SNPs könnten ihre Wirkung beispielsweise ausüben, indem sie einen

Transkriptionsfaktor beeinflussen, der wiederum die Expression verschiedener anderer Gene reguliert.

1.3. Statistisches Testszenario

In eQTL-Studien werden Assoziationen zwischen einem einzelnen SNP und der Expression eines einzelnen Transkripts analysiert. Statistisch gesehen handelt es sich um einen Vergleich der Verteilung einer quantitativen Variablen zwischen drei Gruppen. Das Expressionsniveau des Transkripts ist hierbei die quantitative Größe, während die Genotypen eines SNPs die drei Gruppen bilden. Wenn der SNP einen Einfluss auf die Genexpression hat, also mit der Expression dieses Transkripts assoziiert ist, ist die Expression für jeden Genotyp verschieden. Beispielsweise könnte einer der homozygoten Genotypen eine geringe und der andere homozygote Genotyp eine hohe Expression bewirken, während der heterozygote Genotyp zu einer mittleren Genexpression führt.

Statistisch gesehen können sich die Verteilungen der Expressionswerte in verschiedenen Parametern unterscheiden. Neben Unterschieden in der Lage könnten sich grundsätzlich die Varianzen, die Schiefe oder allgemein die Verteilungsform ändern, woraus sich verschiedene mögliche Nullhypothesen ergeben. Die allgemeinen Hypothesen testen auf die Unterschiedlichkeit der Verteilungen:

$$H_0 : F_0(x) = F_1(x) = F_2(x) \text{ für alle } x \quad (1.1)$$

$$H_1 : F_i(x) \neq F_j(x)$$

für mindestens ein x und ein Paar (i, j) mit $i, j = 0, 1, 2$ und $i \neq j$,

wobei H_0 und H_1 die Null- und Alternativhypothese sowie F_0, F_1 und F_2 beliebige stetige Verteilungsfunktionen der Expressionswerte X für die Genotypen 0, 1 und 2 bezeichnen. Hierbei geben die Werte die Anzahl der seltenen Allele an.

Meistens ist man jedoch nicht an einem allgemeinen Unterschied der Verteilungsformen interessiert, sondern nur an einer Veränderung in der durchschnittlichen

oder mittleren Genexpression bei den drei Genotypen. Also beschränkt man die Hypothesen auf den Vergleich eines Lageparameters $\theta_i \in \mathbb{R}$:

$$\begin{aligned} H_0 & : \theta_0 = \theta_1 = \theta_2 & (1.2) \\ H_1 & : \theta_i \neq \theta_j \\ & \text{für mindestens ein Paar } (i, j) \text{ mit } i, j = 0, 1, 2 \text{ und } i \neq j. \end{aligned}$$

In dieser Situation wird normalerweise vorausgesetzt, dass sich die Verteilungen nur im Parameter θ_i unterscheiden. Dies bedeutet insbesondere, dass gleiche Varianzen in den drei Genotyp-Gruppen angenommen werden. Unter diesen Annahmen lassen sich die Hypothesen auch folgendermaßen darstellen:

$$\begin{aligned} H_0 & : F_0(x) = F_1(x) = F_2(x) & (1.3) \\ H_1 & : F_0(x - \theta_0) = F_1(x - \theta_1) = F_2(x - \theta_2) \\ & \text{für mindestens ein Paar } (i, j) \text{ mit } i, j = 0, 1, 2 \text{ und } \theta_i \neq \theta_j. \end{aligned}$$

Das Standardverfahren für dieses Testszenario ist die Varianzanalyse, die zusätzlich noch Normalverteilung in den drei Gruppen voraussetzt. Nichtparametrische Methoden wie beispielsweise der Kruskal-Wallis-Test basieren auf der allgemeinen Nullhypothese in Gleichung 1.1 und lassen sich nur als einen Test auf einen Lageparameter interpretieren, wenn die Verteilungsformen bis auf die Lage gleich sind, also die Nullhypothese in Gleichung 1.3 angenommen wird.

1.4. Zielsetzung

In den meisten eQTL-Studien wird die Varianzanalyse oder die lineare Regression verwendet. Nur in einzelnen Studien werden zusätzlich oder ausschließlich nichtparametrische Methoden wie der Kruskal-Wallis-Test [82] oder der Spearman-Rangkorrelationskoeffizient [20, 35] genutzt.

Problematisch ist jedoch, dass die Expressionswerte auch nach intensiver Vorverarbeitung inklusive Logarithmierung häufig nicht normalverteilt sind. Stattdes-

sen haben die Verteilungen mehr Verteilungsmasse an den Rändern, oder sie enthalten Ausreißer. Solche Verletzungen der Modellannahmen bei den verschiedenen Verfahren können zu einer erhöhten Anzahl an signifikanten Assoziationen führen, die nicht von anderen statistischen Methoden mit weniger strengen Annahmen gefunden werden [90]. Wahrscheinlich handelt es sich somit um falsch-positive Ergebnisse, die jedoch schwierig zu identifizieren sind, da in eQTL-Studien mehrere Milliarden Tests (z.B. 500.000 SNPs · 10.000 Transkripte = 5 Milliarden Tests) durchgeführt werden.

Wünschenswert ist also für eQTL-Studien ein statistisches Verfahren mit den folgenden Eigenschaften. Zum einen soll es robust sein, d.h. auch bei nicht normalverteilten Daten soll das nominelle Signifikanzniveau eingehalten werden, um die Anzahl falsch-positiver Ergebnisse zu begrenzen. Zum anderen soll das Verfahren eine hohe Güte haben, also wahre Assoziationen mit möglichst hoher Wahrscheinlichkeit entdecken können.

In dieser Arbeit werden anhand von Simulationsstudien verschiedene statistische Methoden bezüglich dieser Kriterien untersucht und optimierte Tests vorgeschlagen. Hierbei werden sowohl klassische statistische Verfahren wie die Varianzanalyse, der Kruskal-Wallis-Test und der Median-Test, aber auch weniger bekannte Methoden verwendet. Zu letzteren gehören zum einen Methoden, die auf der Mutual-Information, einem Konzept aus der Informationstheorie, basieren. Zum anderen handelt es sich um sogenannte adaptive Verfahren, die zunächst Eigenschaften der Verteilungen schätzen, um anschließend einen geeigneten Test auszuwählen.

1.5. Aufbau der Arbeit

In den beiden folgenden Kapiteln werden zwei Simulationsstudien vorgestellt, die sich sowohl in den untersuchten statistischen Methoden als auch in den simulierten Daten unterscheiden. Kapitel 2 vergleicht die klassischen und Mutual-Information-basierten Verfahren und basiert auf der Publikation Szymczak et al. [89]. In dieser Simulation werden vor allem schiefe Verteilungen und solche mit

viel Masse in den Verteilungsrändern simuliert, wobei sich zusätzlich die Variabilität und der Grad der Schiefe in den Genotyp-Gruppen unterscheiden. Die zweite Simulationsstudie in Kapitel 3 konzentriert sich auf die adaptiven Methoden. Für die Simulation wird die g -und- k -Verteilungsfamilie verwendet, bei der Parameter für Schiefe und Verteilungsmasse an den Rändern unabhängig voneinander variiert werden können. In beiden Kapiteln wird jeweils zunächst die Zielsetzung der jeweiligen Studie vorgestellt, bevor die statistischen Methoden sowie die Simulationsstudie beschrieben werden. Anschließend werden die Ergebnisse erläutert und im letzten Abschnitt diskutiert.

Im Kapitel 4 werden ausgewählte Verfahren verwendet, um SNPs zu detektieren, die einen Einfluss auf die Genexpression in Monozyten haben. Die hierfür verwendeten Daten sind Teil der populationsbasierten Gutenberg-Herz-Studie, die das Ziel hat, Risikofaktoren für verschiedene kardiovaskuläre Krankheiten zu identifizieren. Dieses Kapitel beschreibt nach der Zielsetzung zunächst die Daten, bevor näher auf die Methoden eingegangen wird, die für die Vorverarbeitung der Genotyp- und Expressionsdaten verwendet werden. Wie in den beiden vorherigen Kapiteln folgen ein Ergebnis- und ein Diskussionsteil.

Die Arbeit schließt mit einer gemeinsamen Diskussion aller drei Kapitel, die auch einen Ausblick auf weitere Fragestellungen gibt.

2. Simulationsstudie I: Klassische und Mutual-Information-basierte Verfahren

Erschienen in:

Szymczak S, Igl BW, Ziegler A: Detecting SNP-expression associations: A comparison of mutual information and median test with standard statistical approaches. Stat Med 28, 3581-3596 (2009)

2.1. Zielsetzung

In dieser Simulationsstudie wird zunächst untersucht, wie sich Verletzungen der Modellannahmen von klassischen statistischen Verfahren wie der Varianzanalyse und des Kruskal-Wallis-Tests auf die empirischen Fehler 1. Art und die empirische Güte auswirken. Hierfür werden zum einen symmetrische Verteilungen mit unterschiedlich viel Verteilungsmasse an den Rändern und zum anderen schiefe Verteilungen simuliert. Bei letzteren wird zusätzlich die Stärke oder die Richtung der Schiefe zwischen den Genotypen variiert. Die Standardverfahren werden mit robusteren Alternativen wie dem Median-Test und einer kürzlich publizierten Mutual-Information-basierte Methode [92] verglichen. Zusätzlich wird eine weniger rechenintensive Modifikation dieses Verfahrens vorgeschlagen [89].

2.2. Methoden

Für die Beschreibung der Methoden sowie der Simulationsstudie wird die folgende Notation verwendet. Die Expressionswerte werden durch die stetige Zufallsvariable (ZV) X beschrieben, während die diskrete ZV Y den Genotyp eines biallelischen SNPs angibt. Hierbei werden die Werte 0, 1 und 2 für die Genotypen wie in Abschnitt 1.3 verwendet. Die Stichprobengröße beträgt n und die Anzahl Personen pro Genotyp i wird mit n_i bezeichnet. Für den SNP mit der relativen Häufigkeit p des selteneren Allels soll das Hardy-Weinberg-Gleichgewicht gelten, so dass sich die absoluten Genotyp-Häufigkeiten durch $n_i = n \cdot p^{2-i} \cdot (1-p)^i$, $i = 0, 1, 2$ ergeben.

In diesem Abschnitt werden zunächst die Varianzanalyse, der Kruskal-Wallis-Test und der Median-Test als klassische statistische Tests für einen Lageunterschied in drei Gruppen vorgestellt. Anschließend wird zuerst der Hintergrund der Mutual-Information-basierten Methoden beschrieben, bevor auf die aus der Literatur bekannte sowie die modifizierte Teststatistik eingegangen wird.

2.2.1. Klassische Verfahren

Die klassischen statistischen Methoden unterscheiden sich hinsichtlich der Informationen, die in die jeweiligen Teststatistiken eingehen. Während die parametrische Varianzanalyse die vollständigen quantitativen Informationen der Expressionsdaten ausnutzt, werden diese beim nichtparametrischen Ansatz zunächst in Ränge transformiert, auf denen anschließend der Test durchgeführt wird. Beim weniger bekannten Median-Test geht sogar nur noch die Information ein, ob ein Wert größer oder kleiner als der Median der Gesamtstichprobe ist.

Varianzanalyse

Das klassische parametrische Verfahren für einen Test auf einen Lageunterschied in mehr als zwei Gruppen ist die Varianzanalyse, die neben der Normalvertei-

lung auch gleiche Varianzen bei allen Genotypen voraussetzt. Die Teststatistik vergleicht die Varianz zwischen den Genotypen und die Varianz innerhalb der Genotypen. Für die Genotyp-spezifischen Stichproben $\{X_{ij}\}$ mit Genotyp $i = 0, 1, 2$ und $j = 1, \dots, n_i$ ist sie definiert als

$$F = \frac{(n-3) \sum_{i=0}^2 n_i (\bar{X}_i - \bar{X})^2}{2 \sum_{i=0}^2 \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2},$$

wobei \bar{X}_i den empirischen Mittelwert der Expressionswerte von Genotyp i und \bar{X} den Gesamtmittelwert der Expressionswerte bezeichnen. Unter der Nullhypothese gleicher Normalverteilungen und damit gleicher Mittelwerte ist F F-verteilt mit 2 und $n - 2$ Freiheitsgraden (FG).

Wenn die Fallzahlen und die Varianzen in den verschiedenen Gruppen unterschiedlich sind, wird das nominelle Signifikanzniveau nicht eingehalten. Der empirische Fehler 1. Art ist zu hoch, wenn die Varianzen in der kleinsten Gruppe am größten sind, und im umgekehrten Fall ist die Varianzanalyse konservativ [14, 29]. Dies gilt sowohl bei normalverteilten Daten als auch bei Daten mit viel Verteilungsmasse an den Rändern [14].

Da die Annahme gleicher Varianzen in vielen Fällen und vor allem bei Genexpressionsdaten nicht realistisch ist, wird in dieser Arbeit die Modifikation von Welch [98] verwendet. Die folgende Teststatistik berücksichtigt heterogene Varianzen bei den Genotypen, indem die Quadratsummen mit den Genotyp-spezifischen Varianzen gewichtet werden:

$$F_W = \frac{\frac{1}{2} \sum_{i=0}^2 w_i (\bar{X}_i - \tilde{X})^2}{1 + \frac{1}{4} \sum_{i=0}^2 \frac{1}{n_i - 1} \left(1 - \frac{w_i}{w}\right)^2}$$

mit $S_i =$ empirische Standardabweichung, $w_i = \frac{n_i}{S_i^2}$, $w = \sum_{i=0}^2 w_i$ und $\tilde{X} = \sum_{i=0}^2 \frac{w_i}{w} \bar{X}_i$. Unter der Nullhypothese gleicher Mittelwerte bei normalverteilten Expressionswerten ist F_W asymptotisch F-verteilt mit 2 und f FG, wobei $f = \frac{3}{8} \sum_{i=0}^2 \frac{(1 - \frac{w_i}{w})^2}{n_i - 1}$.

Dieser Test hält bei Daten mit ungleichen Varianzen das Signifikanzniveau ein, wobei die Güte bei gleichen Varianzen nur wenig geringer ist als die Güte des F-Tests [12].

Kruskal-Wallis-Test

Die nichtparametrische Variante der Varianzanalyse ist der Kruskal-Wallis-Test. Obwohl hier keine Normalverteilung vorausgesetzt wird, kann diese Methode nur dann als ein Test auf Lageunterschiede interpretiert werden, wenn die Verteilungen bis auf den Lageparameter gleich sind [57]. Diese Voraussetzung enthält also auch die Annahme, dass die Varianzen gleich sind. Die Teststatistik basiert auf Rängen, so dass zunächst jedem Expressionswert X_{ij} der jeweilige Rang R_{ij} in der Gesamtstichprobe X_1, \dots, X_n zugewiesen wird. Anschließend werden die Rangsummen $R_i = \sum_{j=1}^{n_i} R_{ij}$ für jeden Genotyp i berechnet. Die Teststatistik verwendet die gewichteten quadrierten Differenzen der beobachteten und erwarteten Rangsummen:

$$KW = \frac{12}{n(n+1)} \sum_{i=0}^2 \frac{1}{n_i} \left(R_i - \frac{n_i(n+1)}{2} \right)^2.$$

Unter der Nullhypothese gleicher Verteilungsformen ist KW asymptotisch χ^2 -verteilt mit 2 FG.

Bei homoskedastischen, normalverteilten Daten ist der Kruskal-Wallis-Test nur wenig schlechter als die Varianzanalyse, da die asymptotische relative Effizienz 0,955 beträgt [16]. Jedoch hat der Kruskal-Wallis-Test beispielsweise bei schiefen Verteilungen eine höhere Güte [63]. Ungleiche Varianzen führen vor allem bei schiefen Verteilungen zu überhöhten Fehlern 1. Art [26]. Dies gilt jedoch sogar bei Normalverteilung und gleichen Fallzahlen [106]. Zu ähnlichen Ergebnissen auch bei schiefen Verteilungen kommen Fagerland und Sandvik [26], jedoch untersuchten sie nur den für den Zweistichprobenfall äquivalenten Wilcoxon-Mann-Whitney-Test.

Median-Test

Ein weiteres nichtparametrisches Verfahren ist der Median-Test [70]. Die Voraussetzungen für diesen Test sind schwächer als beim Kruskal-Wallis-Test, da die Verteilungen nur die gleiche Dichte in der Nähe des Medians haben müssen [27]. Für die Teststatistik wird zunächst der gemeinsame Median der Gesamtstichprobe bestimmt. Unter der Nullhypothese gleicher Mediane sollte jeweils die Hälfte der Expressionswerte bei jedem Genotyp größer als dieser gemeinsame Median sein. Im nächsten Schritt wird also eine 3x2-Kontingenztabelle erstellt, die für jeden Genotyp (Zeilen) angibt, wie viele Expressionswerte größer oder kleiner als der gemeinsame Median sind (Spalten). Die Teststatistik vergleicht nun die beobachteten und erwarteten Häufigkeiten und entspricht damit der Statistik des χ^2 -Tests:

$$MED = \sum_{i=0}^2 \sum_{j=1}^2 \left(\frac{O_{ij} - E_{ij}}{E_{ij}} \right)^2,$$

wobei O_{ij} und E_{ij} die beobachteten bzw. die erwarteten Häufigkeiten in Zeile (Genotyp) i und Spalte j bezeichnen. Unter der Nullhypothese gleicher Mediane ist MED asymptotisch χ^2 -verteilt mit 2 FG.

Der Median-Test hat einige Vorteile gegenüber rangbasierten Verfahren wie dem Kruskal-Wallis-Test [11]. Er ist sensitiv gegenüber Lokationsunterschieden, ohne stark von unterschiedlichen Verteilungsformen beeinflusst zu werden. Zusätzlich haben unterschiedliche Fallzahlen in den Genotypgruppen keinen besonders starken Effekt. Andererseits führt die ausschließliche Verwendung von Häufigkeiten anstelle der quantitativen Werte oder Ränge zu einem Güteverlust. Beispielsweise beträgt bei Normalverteilung die asymptotische relative Effizienz im Vergleich zur Varianzanalyse ungefähr 0,637 [16]. Dieser Güteverlust zeigt sich auch in einer Simulationsstudie bei verschiedenen anderen symmetrischen Verteilungen [27], so dass die Autoren empfehlen, diesen Test nicht mehr einzusetzen. Die Simulationsstudie hat jedoch einige Nachteile. Sie konzentriert sich auf symmetrische Verteilungen, wobei zusätzlich nur Szenarien mit gleichen Varianzen betrachtet werden. Zu schiefen Verteilungen wird nur in einem kurzen Abschnitt bemerkt, dass auch in diesen Situationen alternative Verfahren dem Median-Test überlegen sind ohne jedoch näher auf die konkreten simulierten Szenarien einzugehen.

2.2.2. Mutual-Information-basierte Verfahren

Einen ähnlichen Ansatz wie der Median-Test verfolgt der sogenannte Quantitative-Mutual-Information-Score, der auf dem Konzept der Mutual-Information aus der Informationstheorie basiert. Mutual-Information ermöglicht eine Aussage über die statistische Abhängigkeit zweier ZV. Im Gegensatz zum Korrelationskoeffizienten nach Pearson, der die lineare Abhängigkeit zweier ZV misst, berücksichtigt Mutual-Information jede mögliche Art von Abhängigkeit, insbesondere auch nichtlineare Zusammenhänge. Von Vorteil ist weiterhin, dass Mutual-Information unempfindlich gegenüber fehlenden Werten oder Ausreißern ist.

Mutual-Information wurde in den letzten Jahren erfolgreich für eine Reihe von Analysen genetischer und genomischer Daten verwendet. Beispielsweise wurden Genexpressionsdaten mit Mutual-Information als Distanzmaß geclustert [75] oder Mutual-Information zur Auswahl von SNPs [90] oder Transkripten [62] für Klassifikationsanalysen eingesetzt. Andere Studien nutzten Mutual-Information, um regulatorische Netzwerke basierend auf Expressionsdaten zu erstellen [5]. Außerdem wurden mithilfe von Mutual-Information genetische Varianten identifiziert, die alleine oder in Interaktionen das Risiko für komplexe Erkrankungen erhöhen [19, 102].

Auch eQTL-Studien wurden unter Verwendung von Mutual-Information durchgeführt. Tsalenko et al. [92] schlugen eine modifizierte Version der Mutual-Information, genannt Quantitative-Mutual-Information-Score (QMIS), als Alternative zur Varianzanalyse vor, um Assoziationen zwischen SNPs und Genexpressionswerten zu untersuchen. Obwohl dieser Score in verschiedenen Studien angewendet wurde [55, 56], sind jedoch bisher keine Informationen über die Validität oder Güte, v.a. im Vergleich mit anderen statistischen Methoden, vorhanden. Ein Vergleich von QMIS und verschiedenen klassischen statistischen Verfahren in einer eQTL-Studie mit weniger als 100 unabhängigen Personen zeigt, dass die methodenspezifischen Signifikanzen stark von der Verteilungssituation abhängen [90].

In den nächsten beiden Unterabschnitten werden zunächst die grundlegenden

Konzepte aus der Informationstheorie, Entropie und die darauf basierende Mutual-Information, eingeführt, bevor QMIS näher beschrieben wird.

Entropie

Die Entropie einer ZV beschreibt den erwarteten Informationsgehalt einer Realisierung der ZV, d.h. sie gibt an, wie überraschend der beobachtete Wert der ZV ist. Für eine diskrete ZV Y ist die Entropie nach Shannon [83] definiert als

$$H(Y) = - \sum_{y \in \mathcal{Y}} p_Y(y) \log_2 p_Y(y).$$

Hierbei bezeichnet $p_Y(y) = P_Y(Y = y)$ die Wahrscheinlichkeitsverteilung von Y , wobei die Realisierungen y aus der Menge aller möglichen Werte \mathcal{Y} stammen. Sie nimmt nur nicht-negative Werte an und je höher die Entropie ist, desto unsicherer sind die Realisierungen der ZV. Die Entropie ist genau dann Null, wenn Y eine degenerierte ZV ist, die nur einen Wert annehmen kann. Im Gegensatz dazu hat eine gleichverteilte ZV maximale Entropie. Der maximale Wert hängt hierbei nur von der Anzahl der möglichen Realisierungen von Y ab.

Mutual-Information

Das Konzept der Mutual-Information erweitert die Entropie auf den Fall zweier diskreter ZV Y und Z und beschreibt, wie viel Information die eine ZV über die andere ZV enthält. Mit anderen Worten, Mutual-Information gibt die wechselseitige Abhängigkeit an, d.h. wie stark verringert sich die Unsicherheit bezüglich Y , wenn Z bekannt ist und umgekehrt. Im diskreten Fall ist Mutual-Information definiert als [18]

$$MI(Y, Z) = \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} p_{YZ}(y, z) \log_2 \frac{p_{YZ}(y, z)}{p_Y(y)p_Z(z)},$$

wobei $p_{YZ}(y, z)$ die gemeinsame Wahrscheinlichkeitsverteilung von Y und Z bezeichnet und die Randverteilungen für Y und Z durch $p_Y(y)$ and $p_Z(z)$ gege-

ben sind. Mutual-Information ist Null bei zwei unabhängigen ZV und maximal, wenn die beiden ZV linear korreliert sind. In diesem Fall enthält die eine ZV die gesamte Information über die andere ZV.

Mutual-Information und Entropie hängen über die folgenden Gleichungen zusammen:

$$MI(Y, Z) = H(Y) - H(Y|Z) = H(Z) - H(Z|Y) = H(Y) + H(Z) - H(Y, Z). \quad (2.1)$$

Hierbei bezeichnen $H(Y|Z)$, $H(Z|Y)$ und $H(Y, Z)$ die bedingte und gemeinsame Entropie.

Mutual-Information kann außerdem als Kullback-Leibler-Distanz zwischen der gemeinsamen Wahrscheinlichkeitsverteilung $p_{YZ}(y, z)$ und dem Produkt der beiden Randverteilungen $p_Y(y) \cdot p_Z(z)$ ausgedrückt werden [18].

Für die Schätzung der Mutual-Information werden die absoluten Häufigkeiten n_{ij} verwendet, wobei hiermit Beobachtungen mit $Y = i$ und $Z = j$ bezeichnet werden. Entsprechend gilt $n_{i.} = \sum_{j \in \mathcal{Z}} n_{ij}$, $n_{.j} = \sum_{i \in \mathcal{Y}} n_{ij}$ und $n = \sum_{i \in \mathcal{Y}} n_{i.} = \sum_{j \in \mathcal{Z}} n_{.j}$. Empirische Häufigkeiten werden geschätzt durch $\hat{p}_Y(i) = \frac{n_{i.}}{n}$, so dass sich die geschätzte Entropie ergibt als

$$\hat{H}(Y) = - \sum_{i \in \mathcal{Y}} \frac{n_{i.}}{n} \log_2 \frac{n_{i.}}{n}.$$

Mutual-Information wird dann unter Verwendung von Gleichung (2.1) geschätzt als

$$\widehat{MI}(Y, Z) = - \sum_{i \in \mathcal{Y}} \frac{n_{i.}}{n} \log_2 \frac{n_{i.}}{n} - \sum_{j \in \mathcal{Z}} \frac{n_{.j}}{n} \log_2 \frac{n_{.j}}{n} + \sum_{i \in \mathcal{Y}} \sum_{j \in \mathcal{Z}} \frac{n_{ij}}{n} \log_2 \frac{n_{ij}}{n}. \quad (2.2)$$

Göbel et al. [30] haben gezeigt, dass die geschätzte Mutual-Information mit der Teststatistik des χ^2 -Tests zusammenhängt:

$$\hat{\chi}^2(Y, Z) = 2n \ln 2 \cdot \widehat{MI}(Y, Z). \quad (2.3)$$

Unter bestimmten Bedingungen lässt sich die Verteilung der geschätzten Mutual-

Information für zwei unabhängige diskrete ZV Y und Z durch eine Gammaverteilung mit den Parametern $\alpha = \frac{1}{2}(|Y| - 1)(|Z| - 1)$ und $\beta = \frac{1}{n \cdot \ln 2}$ approximieren [30], wobei $|Y|$ und $|Z|$ die Anzahl möglicher Realisierungen der ZV Y und Z angeben. In dieser Parametrisierung hat die Gammaverteilung einen Erwartungswert von $\alpha\beta$ und eine Varianz von $\alpha\beta^2$.

Quantitative-Mutual-Information-Score

Eine Erweiterung der Mutual-Information auf den Fall einer diskreten ZV Y und einer stetigen ZV X , wie z.B. Genexpressionswerte, ist möglich. Sie ist in dieser Situation definiert als

$$MI(X, Y) = \sum_{y \in \mathcal{Y}} \int_{S_x} f_{XY}(x, y) \log_2 \frac{f_{XY}(x, y)}{f_X(x)p_Y(y)} dx,$$

wobei S_x den Träger und $f_X(x)$ die Dichtefunktion der quantitativen ZV X bezeichnen [19]. Entsprechend gibt $f_{XY}(x, y)$ die gemeinsame Dichtefunktion der ZV X und Y an. Die Dichten können mithilfe von numerischer Integration geschätzt werden. Jedoch müssen bei dieser rechenintensiven Methode die Aspekte der numerischen Stabilität und Genauigkeit beachtet werden [19].

Alternativ kann die quantitative ZV X diskretisiert werden, beispielsweise durch eine Dichotomisierung. Für einen Grenzwert q wird eine Variable $X_q = I(X < q)$ definiert, wobei $I(\cdot)$ die Indikatorfunktion bezeichnet. X_q gibt also an, ob der Expressionswert X kleiner oder größer als der Grenzwert q ist.

Für die Wahl des Grenzwerts q werden in dieser Arbeit zwei verschiedene Ansätze betrachtet. Tsalenko et al. [92] verwenden für ihren Quantitative-Mutual-Information-Score (QMIS) die maximale Mutual-Information über alle möglichen Grenzwerte:

$$QMIS_{max}(X, Y) = \max_{q \in \mathbb{R}} MI(X_q, Y).$$

Eine Alternative ist die Verwendung des globalen Medians der Genexpressionsdaten X als Grenzwert q :

$$QMIS_{med}(X, Y) = MI(X_{med}, Y),$$

wobei med den Median der Gesamtstichprobe X_1, \dots, X_n angibt [89].

Der QMIS für die Genexpression X und den Genotyp Y wird unter Verwendung von Gleichung (2.2) geschätzt als

$$\widehat{MI}(X_q, Y) = - \sum_{i \in \mathcal{Y}} \frac{n_i}{n} \log_2 \frac{n_i}{n} - \sum_{j=1}^2 \frac{n_{.j}}{n} \log_2 \frac{n_{.j}}{n} + \sum_{i \in \mathcal{Y}} \sum_{j=1}^2 \frac{n_{ij}}{n} \log_2 \frac{n_{ij}}{n}.$$

Diese Schätzung und Gleichung (2.3) zeigen die Ähnlichkeit von $QMIS_{med}$ und dem Median-Test, da beide den globalen Median der Expressionswerte verwenden, um die quantitativen Expressionswerte zu diskretisieren.

2.3. Simulationen

In diesem Abschnitt wird als erstes auf das allgemeine Simulationsszenario eingegangen, das dieser und auch der im Kapitel 3 beschriebenen Simulationsstudie zugrunde liegt. Hierbei wird die interessierende Nullhypothese vorgestellt und die Wahl des passenden Lageparameters diskutiert. Anschließend werden die simulierten Verteilungen beschrieben und abschließend allgemeine Parameter der Simulationsstudie genannt.

2.3.1. Simulationsszenario

Für diese Arbeit ist die parameterbasierte Nullhypothese (Gleichung 1.2) von Interesse. Es ist also ein Unterschied in der mittleren oder durchschnittlichen Expression zwischen den drei Genotypen relevant. Jedoch wird hier nicht die zusätzliche Annahme gemacht, dass die Verteilungsformen auch unter der Alter-

nativhypothese bis auf die Lage gleich sein sollen. Stattdessen sollen Lageunterschiede unabhängig von der Verteilungsform entdeckt werden, d.h. es soll auch in solchen Fällen ein Unterschied gefunden werden, wenn die Verteilungen bezüglich anderer Parameter wie Varianz oder Schiefe unterschiedlich sind. Mit anderen Worten, es wird auf Gleichheit der Lage unabhängig von den vorliegenden Verteilungen getestet.

Als Lageparameter können beispielsweise der Mittelwert oder der Median verwendet werden. Bei symmetrischen Verteilungen sind Mittelwert und Median identisch, jedoch ist der Mittelwert nur dann optimal, wenn die Verteilungen wenig Masse an den Rändern haben und keine Ausreißer enthalten [59]. Im Fall von Verteilungen mit viel Masse an den Rändern ist der Median effizienter [8, 84] und somit zu bevorzugen, ebenso wie bei vorhandenen Ausreißern, da die Schätzung des Mittelwerts ansonsten stark beeinflusst wird. Bei schiefen Verteilungen ist ebenfalls der Median besser geeignet, da er im Allgemeinen näher an der „Hauptmasse“ der Daten liegt und somit eher einem „typischen“ Wert der Verteilung entspricht [z. B. 21, 100]. Besonders relevant ist die Wahl des Lageparameters, wenn Verteilungen mit unterschiedlicher Form verglichen werden sollen. Beispielsweise könnten im Fall einer symmetrischen und einer schiefen Verteilung die Mittelwerte gleich, aber die Mediane unterschiedlich sein oder umgekehrt. Die Schlussfolgerung, dass ein Lageunterschied und damit eine Assoziation vorliegt, hängt also von dem verwendeten Parameter ab. Da Expressionsdaten häufig nicht symmetrisch verteilt sind und auch Ausreißer enthalten können, wird in dieser Arbeit der Median als Lageparameter verwendet.

Unter der Alternativhypothese wird ein additives genetisches Modell simuliert, wobei die Expression des heterozygoten Genotyps zwischen der Expression der beiden homozygoten Genotypen liegt. Dies stellt jedoch keine Einschränkung dar, da mit den in dieser Arbeit verwendeten statistischen Methoden ein allgemeiner Test auf einen Lageunterschied zwischen den drei Genotypen durchgeführt wird. In anderen Worten, mit diesen Verfahren ist es ebenfalls möglich, ein dominantes oder rezessives Modell zu identifizieren, bei dem ein bzw. zwei Risikoallele zu einer anderen mittleren Expression führen. Auch ein Unterschied zwischen dem heterozygoten und den beiden homozygoten Genotyp (Heterozygoten-Vorteil) kann detektiert werden.

2.3.2. Simulierte Verteilungen

Neben schiefen und symmetrischen Verteilungen wird in dieser Simulationsstudie in den meisten Situationen zusätzlich die Varianz in den drei Gruppen unterschiedlich gewählt. Bei den symmetrischen Verteilungen wird außerdem die Masse an den Verteilungsrändern variiert. Die letztgenannte Eigenschaft wird im Englischen mit „tail length“ oder „tail weight“ bezeichnet und verschiedene Verteilungen als „short-“, „medium-“ oder „long-tailed“ klassifiziert [25]. Anstelle des oben erwähnten etwas sperrigen Begriffs werden im weiteren Verlauf dieser Arbeit analog zum Englischen die Formulierungen kurze bzw. lange Verteilungsschwänze verwendet.

Schiefe Verteilungen

Für die Simulation asymmetrisch verteilter Expressionsdaten wird die χ^2 -Verteilung genutzt, deren Varianz und Schiefe über die Anzahl der Freiheitsgrade modifiziert werden kann. Jedoch wird auch der Median durch die Freiheitsgrade bestimmt, so dass die simulierten Daten verschoben werden müssen, um die gewünschten Mediane unter der Null- und Alternativhypothese zu erhalten. Es werden drei verschiedene Szenarien schiefer Verteilungen betrachtet, die in Abbildung 2.1 dargestellt sind.

- Szenario Ia (gleich starke Rechtsschiefe)
 χ^2 -Verteilungen mit gleichen Freiheitsgraden ($FG = 1, 2$ oder 3) in allen drei Genotyp-Gruppen
- Szenario Ib (unterschiedlich starke Rechtsschiefe)
 χ^2 -Verteilungen mit unterschiedlichen Freiheitsgraden, wobei die Schiefe sowohl mit der Fallzahl geringer ($FG = 1, 2, 3$; $FG = 1, 3, 5$) als auch größer wird ($FG = 3, 2, 1$; $FG = 5, 3, 1$)

- Szenario Ic (unterschiedliche Richtung der Schiefe)
rechtsschiefe, symmetrische und linksschiefe Verteilungen mit gleicher Varianz ($Var = 2; 4$ oder 6) in allen drei Genotyp-Gruppen

Beim letzten Szenario wird die rechtsschiefe Verteilung wiederum als χ^2 -Verteilung simuliert, während die Expressionswerte für den Genotyp 1 als normalverteilt simuliert werden. Die linksschiefe Verteilung ist eine „negative“ χ^2 -Verteilung $\chi^2\text{-neg}$, die durch eine Veränderung des Vorzeichens der Standard χ^2 -Verteilung generiert wird.

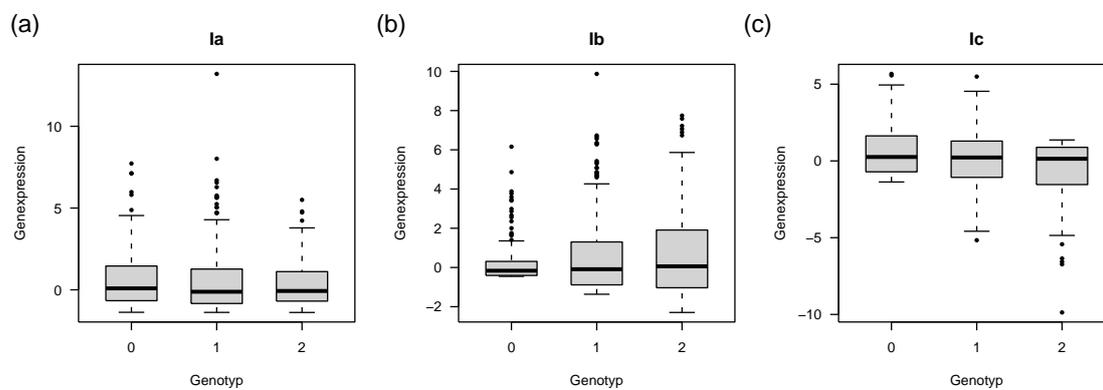


Abbildung 2.1: Beispiele für die verschiedenen Simulationsszenarien schiefer Verteilungen unter der Nullhypothese gleicher Mediane: (a) Szenario Ia: gleich starke Rechtsschiefe, (b) Szenario Ib: unterschiedlich starke Rechtsschiefe, (c) Szenario Ic: Rechtsschiefe, Symmetrie, Linksschiefe. Boxplots geben Median, Quartile, größten Nicht-Ausreißer und die Extrema an.

Symmetrische Verteilungen

Im Szenario II werden Expressionsdaten mit langen Verteilungsschwänzen unter Verwendung der Laplace-Verteilung simuliert, um den Fall von häufig beobachteten Ausreißern zu berücksichtigen. Sowohl gleiche ($Var = 6, 6, 6$) als auch unterschiedliche Varianzen ($Var = 2, 6, 10$; $Var = 10, 6, 2$) werden für die Genotyp-Gruppen verwendet. Die gleichen Parametereinstellungen werden auch für die Simulation normalverteilter Genexpressionsdaten genutzt.

2.3.3. Parameter

Für alle simulierten Szenarien werden die folgenden Parameter verwendet. Für den SNP wird eine relative Häufigkeit des selteneren Allels von $p = 0,5$ oder $p = 0,3$ angenommen. Für eine Gesamtfallzahl von $n = 500$ ergeben sich also die absoluten Genotyp-Häufigkeiten $n_0 = n_2 = 125$ und $n_1 = 250$ für $p = 0,5$ and $n_0 = 245$, $n_1 = 210$ and $n_2 = 45$ für $p = 0,3$. Unter der Nullhypothese werden unabhängig von der Verteilungsform gleiche theoretische Mediane simuliert. Für die Alternativhypothese wird ein additives Modell definiert, wobei die theoretischen Mediane auf 1, 1,25 and 1,5 für die Genotypen 0, 1 und 2 gesetzt werden. Für jedes Simulationsszenario werden 1000 Wiederholungen durchgeführt. Für jedes der simulierten Szenarien werden die p-Werte der Varianzanalyse, des Kruskal-Wallis-Test, des Median-Test und der beiden Mutual-Information-basierten Verfahren $QMIS_{max}$ und $QMIS_{med}$ bestimmt. Für die klassischen Verfahren werden zur besseren Vergleichbarkeit auch empirische p-Werte basierend auf $N = 10^5$ Permutationen dargestellt. Für die Beurteilung der Robustheit eines Tests werden die liberalen Kriterien von Bradley verwendet, die einen Test als robust ansehen, wenn der geschätzte Fehler 1. Art zwischen $0,5 \cdot \alpha$ und $1,5 \cdot \alpha$ liegt [10], wobei hier $\alpha = 0,05$ gesetzt wird.

2.4. Ergebnisse

Für den Vergleich der verschiedenen statistischen Verfahren werden zunächst die empirischen Fehler 1. Art unter der Nullhypothese gleicher Median dargestellt, um die Validität der Methoden zu überprüfen. Anschließend wird für die robusten Tests ein Vergleich der empirischen Güte durchgeführt.

2.4.1. Empirische Fehler 1. Art

Tabelle 2.1 gibt die empirischen Fehler 1. Art für die verschiedenen Simulationsszenarien an. Die empirischen Fehler sind für permutationsbasierte und asymp-

totische p-Werte der Varianzanalyse und des Kruskal-Wallis-Tests vergleichbar, so dass im Folgenden keine entsprechende Unterscheidung gemacht wird. Beim Simulationsszenario Ia mit homoskedastischen χ^2 -verteilten Genexpressionsdaten sind alle untersuchten Methoden robust bezüglich der Bradley-Kriterien. Hierbei sind die Varianzanalyse, der Kruskal-Wallis und der Median-Test geringfügig liberal, während beide QMIS-Ansätze etwas konservativ sind. Im Fall unterschiedlicher Varianzen rechtsschiefer Verteilungen (Szenario Ib) unterscheiden sich die statistischen Tests jedoch deutlich. Kruskal-Wallis-Test und $QMIS_{max}$ haben extrem hohe Fehler 1. Art von ungefähr 50 % und 100 % bei einem nominellen Signifikanzniveau von 5 %. Wie im vorherigen Szenario ist die Varianzanalyse etwas liberal. Die einzigen robusten Verfahren in dieser Situation sind der $QMIS_{med}$ und der Median-Test, wobei letztere Methode aber leicht erhöhte empirische Fehler hat. Wenn wie im Szenario Ic die Richtung der Schiefe unterschiedlich ist, überschreiten die Varianzanalyse, der Kruskal-Wallis-Test sowie $QMIS_{max}$ das nominelle Signifikanzniveau sehr stark, während nur die Methoden $QMIS_{med}$ und Median-Test robust sind.

Die Simulationen symmetrischer Verteilungen (Szenarien II und III) führen zu anderen Ergebnissen. Im Fall von Verteilungen mit langen Schwänzen sowie von Normalverteilungen hält neben $QMIS_{med}$ und Median-Test auch die Varianzanalyse das Signifikanzniveau ein. Im Gegensatz dazu hat der Kruskal-Wallis-Test einen geringfügig erhöhten Fehler 1. Art, wenn der homozygot seltene Genotyp die größte Variabilität in den Genexpressionsdaten hat. $QMIS_{max}$ hat einen extrem erhöhten empirischen Fehler von bis zu 100 % bei unterschiedlich starken Varianzen, ist jedoch robust bei gleichen Varianzen.

2.4.2. Empirische Güte

Tabelle 2.2 zeigt die empirische Güte für alle statistischen Methoden und Simulationsszenarien. Hierbei ist die Güte von nicht robusten Tests in einer kleineren Schrift mit zusätzlichen Klammern dargestellt, und die entsprechenden Verfahren werden nicht in den Gütevergleich miteinbezogen. Wie schon unter der Nullhypothese ist die Güte basierend auf asymptotischen und permutationsbasierten p-Werten bei der Varianzanalyse und dem Kruskal-Wallis-Test vergleichbar.

Im Simulationsszenario Ia sind alle Tests robust basierend auf den Bradley-Kriterien, wobei $QMIS_{max}$ am besten geeignet ist, den vorhandenen Unterschied zu detektieren. Im Gegensatz dazu ist bei den Szenarien Ib and Ic nur ein Vergleich zwischen $QMIS_{med}$ und Median-Test möglich. Hier hat der Median-Test immer eine höhere Güte als $QMIS_{med}$.

Bei symmetrischen Verteilungen (Szenarien II und III) sind nur die Varianzanalyse, $QMIS_{med}$ und Median-Test anwendbar. Interessanterweise hat der Median-Test bei den Laplace-Verteilungen die höchste Güte, während, wie erwartet, die Varianzanalyse bei den Normalverteilungen am besten abschneidet.

2.5. Diskussion

Die Simulationsstudie zeigt, dass nur $QMIS_{med}$ und Median-Test in allen betrachteten Verteilungssituationen das Fehlerniveau einhalten, während die Standardverfahren wie die Varianzanalyse und der Kruskal-Wallis-Test sowie die alternative Methode $QMIS_{max}$ in mindestens einer Situation zu einer bisweilen extrem erhöhten Anzahl falsch positiver Ergebnisse führen.

Es gibt jedoch auch einige Studien, die zeigen, dass der Median-Test in Situationen mit extrem großen Varianzunterschieden nicht das Signifikanzniveau einhält, sondern zu hohe empirische Fehler 1. Art aufweist [74]. Dies tritt vor allem in Situationen auf, in denen die maximale Standardabweichung größer als 5 ist [14]. Jedoch sind solche Varianzen bei Genexpressionsdaten nicht zu erwarten, da die Expressionswerte aller exprimierten Gene zusammen meistens eine geringe Spannweite haben. Der Median-Test ist gegenüber $QMIS_{med}$ vorzuziehen, da er in allen Szenarien und Fällen eine höhere Güte hat. Die ähnliche, aber komplizierter herzuleitende Teststatistik des $QMIS_{med}$ hat also keine Vorteile gegenüber dem in der Statistik eher bekannten Verfahren des Median-Tests. Die Ergebnisse der Simulationsstudie zeigen jedoch die enge Verwandtschaft der beiden Teststatistiken. Im Gegensatz dazu reagiert die ursprüngliche Mutual-Information-basierte Methode $QMIS_{max}$ sowohl auf Lage- als auch auf Varianzunterschiede

und führt somit bei Verteilungen mit Varianzunterschieden zu extrem erhöhten empirischen Fehlern 1. Art.

Die klassischen statistischen Verfahren wie die Varianzanalysen und der Kruskal-Wallis-Test versagen vor allem bei schiefen Verteilungen mit unterschiedlichen Varianzen und führen in diesen Situationen zum Teil zu extrem großen Fehlern 1. Art. Die Erklärung liegt in der hier betrachteten Nullhypothese begründet, die vor allem bei schiefen Verteilungen nicht mit der methodenspezifischen Nullhypothese übereinstimmt. Bei der Varianzanalyse werden unter H_0 gleiche Mittelwerte bei Normalverteilung vorausgesetzt, jedoch sind bei schiefen Verteilungen Mittelwerte und Mediane unterschiedlich. Besonders im Fall von Verteilungen mit Schiefe in verschiedenen Richtungen sind unter der hier betrachteten Nullhypothese die Mittelwerte verschieden, so dass die Varianzanalyse zu Recht zu einer Ablehnung führt. Der Kruskal-Wallis-Test setzt gleiche Verteilungsformen voraus, so dass sich die Nullhypothese als Lokationsunterschied interpretieren lässt. Unterschiedliche Verteilungsformen führen jedoch zu einer ungleichmäßigen Verteilung der Ränge auf die Genotyp-Gruppen, so dass die Teststatistik des Kruskal-Wallis-Test groß wird und damit zur Ablehnung der Nullhypothese führt. Es kann also bei der Anwendung von Standardverfahren dazu kommen, dass fälschlicherweise eine Assoziation zwischen einem SNP und der Expression eines Transkripts angenommen wird, die jedoch nur durch verletzte Modellannahmen zustande gekommen ist.

Tabelle 2.1: Empirische Fehler 1. Art bei einem nominellen Signifikanzniveau $\alpha = 0.05$. Die verschiedenen Simulationsszenarien sind im Text beschrieben. In der ersten Spalte sind für jedes Szenario die verwendeten Verteilungen der Genexpression für die 3 Genotypen angegeben, wobei χ^2 die χ^2 -Verteilung, $\chi^2\text{-neg}$ die „negative“ χ^2 -Verteilung und NV die Normalverteilung bezeichnen. Die Häufigkeit des selteneren Allels wird mit p bezeichnet, und Var gibt die simulierten Varianzen an. Die statistischen Methoden sind die Varianzanalyse (VA), der Kruskal-Wallis-Test (KW), die beiden QMIS-Verfahren (QMIS_{max}, QMIS_{med}) und der Median-Test (M).

Verteilung	p	Var	VA	VA (perm)	KW	KW (perm)	QMIS _{max}	QMIS _{med}	M	
Ia										
χ^2, χ^2, χ^2	0,5	2, 2, 2	0,044	0,037	0,054	0,055	0,033	0,034	0,057	
		4, 4, 4	0,052	0,053	0,054	0,051	0,028	0,035	0,054	
		6, 6, 6	0,050	0,049	0,049	0,048	0,026	0,032	0,047	
	0,3	2, 2, 2	0,072	0,056	0,060	0,061	0,035	0,042	0,062	
		4, 4, 4	0,053	0,043	0,037	0,037	0,037	0,048	0,059	
		6, 6, 6	0,056	0,051	0,057	0,058	0,048	0,042	0,055	
Ib										
χ^2, χ^2, χ^2	0,5	2, 4, 6	0,072	0,069	0,413	0,409	1,000	0,031	0,062	
		6, 4, 2	0,063	0,055	0,410	0,405	1,000	0,035	0,053	
		2, 6, 10	0,074	0,069	0,392	0,398	1,000	0,046	0,073	
		10, 6, 2	0,057	0,057	0,378	0,382	1,000	0,040	0,074	
	0,3	2, 4, 6	0,056	0,046	0,520	0,516	1,000	0,039	0,053	
		6, 4, 2	0,106	0,088	0,220	0,218	1,000	0,045	0,052	
		2, 6, 10	0,058	0,048	0,513	0,518	1,000	0,048	0,059	
		10, 6, 2	0,091	0,084	0,153	0,150	1,000	0,050	0,070	
Ic										
$\chi^2, NV, \chi^2\text{-neg}$	0,5	4, 4, 4	0,999	0,999	0,777	0,776	1,000	0,034	0,058	
		6, 6, 6	0,977	0,978	0,560	0,558	1,000	0,042	0,070	
		8, 8, 8	0,920	0,921	0,423	0,421	1,000	0,041	0,056	
		0,3	4, 4, 4	0,998	0,997	0,714	0,771	1,000	0,031	0,038
			6, 6, 6	0,948	0,946	0,479	0,541	1,000	0,040	0,052
			8, 8, 8	0,844	0,842	0,372	0,421	1,000	0,042	0,050
	$\chi^2\text{-neg}, NV, \chi^2$	0,5	4, 4, 4	0,999	0,999	0,770	0,713	1,000	0,038	0,055
			6, 6, 6	0,978	0,979	0,540	0,483	1,000	0,041	0,064
			8, 8, 8	0,917	0,917	0,428	0,368	1,000	0,032	0,055
		0,3	4, 4, 4	0,991	0,990	0,701	0,705	1,000	0,051	0,064
			6, 6, 6	0,923	0,924	0,472	0,476	1,000	0,046	0,052
			8, 8, 8	0,848	0,848	0,352	0,345	1,000	0,045	0,059
II										
Laplace	0,5	6, 6, 6	0,061	0,062	0,055	0,053	0,042	0,031	0,047	
		2, 6, 10	0,058	0,058	0,054	0,058	1,000	0,040	0,055	
	0,3	6, 6, 6	0,046	0,049	0,048	0,047	0,039	0,026	0,047	
		2, 6, 10	0,046	0,045	0,088	0,086	1,000	0,040	0,060	
		10, 6, 2	0,043	0,043	0,025	0,027	0,651	0,034	0,052	
III										
NV	0,5	6, 6, 6	0,058	0,051	0,048	0,047	0,033	0,032	0,052	
		2, 6, 10	0,058	0,059	0,064	0,063	1,000	0,040	0,060	
	0,3	6, 6, 6	0,053	0,046	0,047	0,046	0,050	0,036	0,052	
		2, 6, 10	0,047	0,046	0,084	0,083	1,000	0,045	0,061	
		10, 6, 2	0,047	0,046	0,021	0,023	1,000	0,038	0,055	

Tabelle 2.2: Empirische Güte bei einem nominellen Signifikanzniveau $\alpha = 0.05$. Die verschiedenen Simulationsszenarien sind im Text beschrieben. In der ersten Spalte sind für jedes Szenario die verwendeten Verteilungen der Genexpression für die 3 Genotypen angegeben, wobei χ^2 die χ^2 -Verteilung, $\chi^2\text{-neg}$ die „negative“ χ^2 -Verteilung und NV die Normalverteilung bezeichnen. Die Häufigkeit des selteneren Allels wird mit p bezeichnet, und Var gibt die simulierten Varianzen an. Die statistischen Methoden sind die Varianzanalyse (VA), der Kruskal-Wallis-Test (KW), die beiden QMIS-Verfahren (QMIS_{max}, QMIS_{med}) und der Median-Test (M). Die Güte von nicht robusten Methoden ist in Klammern und einer kleineren Schriftgröße dargestellt.

Verteilung	p	Var	VA	VA (perm)	KW	KW (perm)	QMIS _{max}	QMIS _{med}	M	
Ia										
χ^2, χ^2, χ^2	0,5	2,2,2	0,684	0,668	1,000	1,000	1,000	0,789	0,921	
		4,4,4	0,429	0,420	0,810	0,809	1,000	0,230	0,368	
		6,6,6	0,294	0,291	0,464	0,468	0,753	0,123	0,228	
	0,3	2,2,2	0,588	0,517	1,000	1,000	1,000	0,899	0,937	
		4,4,4	0,311	0,280	0,728	0,729	1,000	0,275	0,357	
		6,6,6	0,223	0,206	0,403	0,403	0,707	0,172	0,206	
Ib										
χ^2, χ^2, χ^2	0,5	2,4,6	0,588	0,589	(0,097)	(0,101)	(1,000)	0,403	0,502	
		6,4,2	0,322	0,316	(0,987)	(0,987)	(1,000)	0,381	0,607	
		2,6,10	0,518	0,510	(0,078)	(0,081)	(1,000)	0,305	0,388	
		10,6,2	0,226	0,222	(0,935)	(0,938)	(1,000)	0,266	0,494	
	0,3	2,4,6	0,500	0,451	(0,130)	(0,129)	(1,000)	0,355	0,431	
		6,4,2	(0,253)	(0,229)	(0,926)	(0,923)	(1,000)	0,611	0,652	
		2,6,10	0,428	0,392	(0,117)	(0,114)	(1,000)	0,232	0,294	
		10,6,2	(0,190)	(0,176)	(0,768)	(0,767)	(1,000)	0,518	0,556	
Ic										
$\chi^2, NV, \chi^2\text{-neg}$	0,5	4,4,4	(0,755)	(0,757)	(0,090)	(0,088)	(1,000)	0,185	0,314	
		6,6,6	(0,601)	(0,605)	(0,081)	(0,086)	(1,000)	0,129	0,240	
		8,8,8	(0,488)	(0,488)	(0,075)	(0,074)	(1,000)	0,096	0,170	
		0,3	4,4,4	(0,665)	(0,657)	(0,078)	(0,077)	(1,000)	0,209	0,267
			6,6,6	(0,490)	(0,487)	(0,067)	(0,067)	(1,000)	0,145	0,188
			8,8,8	(0,392)	(0,385)	(0,079)	(0,082)	(1,000)	0,111	0,157
	$\chi^2\text{-neg}, NV, \chi^2$	0,5	4,4,4	(1,000)	(1,000)	(0,999)	(0,999)	(1,000)	0,273	0,439
			6,6,6	(1,000)	(1,000)	(0,964)	(0,966)	(1,000)	0,136	0,256
			8,8,8	(0,998)	(0,998)	(0,888)	(0,887)	(1,000)	0,106	0,200
		0,3	4,4,4	(1,000)	(1,000)	(0,994)	(0,994)	(1,000)	0,299	0,386
			6,6,6	(0,999)	(0,999)	(0,924)	(0,925)	(1,000)	0,174	0,229
			8,8,8	(0,992)	(0,992)	(0,818)	(0,820)	(1,000)	0,117	0,157
II										
Laplace	0,5	6,6,6	0,280	0,283	0,394	0,396	0,183	0,270	0,453	
		2,6,10	0,360	0,360	0,464	0,456	(1,000)	0,478	0,585	
	0,3	6,6,6	0,253	0,253	0,342	0,339	0,187	0,303	0,399	
		2,6,10	0,287	0,297	(0,469)	(0,472)	(1,000)	0,396	0,509	
		10,6,2	0,325	0,320	0,298	0,296	(0,784)	0,414	0,517	
III										
NV	0,5	6,6,6	0,304	0,311	0,277	0,276	0,100	0,102	0,192	
		2,6,10	0,377	0,372	0,300	0,301	(1,000)	0,194	0,264	
	0,3	6,6,6	0,258	0,259	0,241	0,243	0,131	0,131	0,172	
		2,6,10	0,279	0,276	(0,333)	(0,337)	(1,000)	0,167	0,221	
		10,6,2	0,326	0,323	0,159	0,162	(1,000)	0,188	0,233	

3. Simulationsstudie II: Adaptive Verfahren

3.1. Zielsetzung

Das vorherige Kapitel zeigt, dass der Median-Test in Situationen, die bei eQTL-Studien auftreten können, das Signifikanzniveau einhält. Diese Robustheit führt jedoch zu einem Güteverlust, der besonders ins Gewicht fällt, wenn die Modellannahmen für klassische statistische Verfahren erfüllt sind. Deshalb kann der Median-Test nicht allgemein empfohlen werden, sondern sollte nur in Situationen verwendet werden, in denen die klassischen Verfahren nicht robust sind. Häufig wird eine solche Entscheidung anhand einer grafischen Darstellung der Daten manuell getroffen. Diese Vorgehensweise ist bei einer eQTL-Studie jedoch aufgrund der großen Anzahl an Tests nicht praktikabel. Alternativ können adaptive Methoden eingesetzt werden, die zunächst die vorliegende Datensituation analysieren und anschließend einen geeigneten Test auswählen. Im ersten Schritt werden bei diesen Verfahren sogenannte Selektorstatistiken verwendet, die Eigenschaften der Verteilungen wie Schiefe und Länge der Verteilungsschwänze schätzen. Im Gegensatz zu Vorschalttests, die beispielsweise auf Normalverteilung oder Gleichheit der Varianzen testen, wird also die Entscheidung für einen bestimmten Test nicht auf Basis eines p-Werts, sondern anhand einer von den einzelnen Tests unabhängigen Statistik getroffen. Deshalb kann gezeigt werden, dass adaptive Tests das Signifikanzniveau einhalten Hogg [40]. Im Gegensatz dazu führt ein t-Test oder eine Varianzanalyse mit Vorschalttest zu erhöhten Fehlern 1. Art [99].

In diesem Kapitel werden zunächst zwei aus der Literatur bekannte adaptive lineare Rangtests in einer Simulationsstudie untersucht. Basierend auf den Ergebnissen wird anschließend ein neues Klassifikationsschema für die Auswahl der Tests entwickelt. Zusätzlich wird eine Kombination dieses neuen Verfahrens mit dem Median-Test vorgeschlagen. Für die entsprechenden Simulationen wird eine spezielle Verteilungsfamilie verwendet, bei der die Parameter für die Schiefe und die Länge der Verteilungsschwänze unabhängig voneinander variiert werden können. Mithilfe dieser Familie werden Szenarien mit gleich schiefen oder gleich langen Verteilungen simuliert. Im zweiten Teil dieser Studie werden die bekannten und die neuen adaptiven Tests miteinander verglichen, wobei zusätzlich Szenarien verwendet werden, in denen sich die Genotyp-spezifischen Verteilungen hinsichtlich Schiefe oder Länge der Verteilungsschwänze unterscheiden.

3.2. Methoden

Adaptive Tests verwenden Informationen aus den vorliegenden Daten, um einen passenden Test auszuwählen. Im ersten Schritt werden Eigenschaften der Verteilungen wie Schiefe und Länge der Verteilungsschwänze geschätzt. Anhand dieser sogenannten Selektorstatistiken wird im zweiten Schritt ein für die konkrete Datensituation geeigneter Test ausgewählt und durchgeführt. Dieses Konzept wurde von Hogg et al. [41] ursprünglich für das Lokationsproblem bei 2 Stichproben vorgeschlagen. Durch die Verwendung von rangbasierten Selektorstatistiken und davon unabhängigen linearen Rangstatistiken hält dieser adaptive Test das vorgegebene Signifikanzniveau ein. Mittlerweile wurde dieser Ansatz für weitere statistische Fragestellungen verwendet. Neben Erweiterungen auf den allgemeinen c Stichprobenfall [12] wurden beispielsweise adaptive Tests für Trend- oder Umbrella-Alternativen vorgeschlagen [53] und das Konzept auf Regressionsanalysen übertragen [42]. Zusätzlich wurden adaptive Tests unter Verwendung von U-Statistiken entwickelt [52, 54]. Jedoch ist der asymptotische Gütegewinn dieser Tests im Vergleich zu linearen Rangstatistiken nur gering, während die Durchführung rechenintensiver ist.

Als Alternative zu der an die Daten angepassten Teststatistik werden beim Ma-

ximum-Test [72] verschiedene Teststatistiken berechnet und die Statistik mit dem größten Wert ausgewählt. Die entsprechenden p-Werte für diese maximale Teststatistik werden durch Permutationen bestimmt, wodurch sich die Rechenzeit erheblich verlängert.

Aufgrund der genannten Nachteile der U-Statistiken und des Max-Tests werden in dieser Arbeit zwei adaptive lineare Rangtests basierend auf dem ursprünglichen Konzept untersucht und weiterentwickelt. Da das Verfahren von Hogg direkt auf den c Stichprobenfall übertragbar ist, wird hier auf eine Darstellung des ursprünglichen Verfahrens verzichtet und die Erweiterungen von Büning [12] und Beier [7] beschrieben. Zunächst werden jedoch die einzelnen Komponenten, die Selektorstatistiken sowie die linearen Rangtests dargestellt. In den folgenden Abschnitten wird die gleiche Notation wie in Kapitel 2 verwendet.

3.2.1. Selektorstatistiken

Selektorstatistiken beschreiben Eigenschaften von Verteilungen wie Schiefe, Wölbung (auch Kurtosis oder Exzess genannt), Spitzigkeit oder Länge der Verteilungsschwänze. Es gibt verschiedene Methoden, solche Selektorstatistiken zu definieren und zu schätzen. Die bekanntesten Maßzahlen für Schiefe und Wölbung basieren auf Momenten. Die Schiefe β_1 einer Zufallsvariablen X mit Verteilungsfunktion F ist definiert als

$$\beta_1 = \frac{\mu_3}{\sigma^3},$$

wobei $\mu_3 = E((X - E(X))^3)$ das 3. zentrale Moment und σ die Standardabweichung bezeichnen. Für eine symmetrische Verteilung gilt $\beta_1 = 0$ und eine Verteilung mit $\beta_1 > 0$ bzw. $\beta_1 < 0$ heißt rechts- bzw. linksschief.

Der Koeffizient der Wölbung β_2 ist definiert als

$$\beta_2 = \frac{\mu_4}{\sigma^4}.$$

Da die Normalverteilung für β_2 den Wert 3 annimmt, gibt es einen alternativen Schätzer für die Wölbung:

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3.$$

In der Literatur gibt es verschiedene Ansätze, um den Zusammenhang zwischen der Wölbung und der Form einer Verteilung zu beschreiben [104]. Wölbung kann als eine Umverteilung der Wahrscheinlichkeitsmasse von den „Schultern“ einer Verteilung ins Zentrum und in die Enden der Verteilung interpretiert werden. Also ist sie eine Kombination aus Spitzigkeit und Länge der Schwänze einer Verteilung. Problematisch ist jedoch, dass Verteilungen mit dem gleichen Wert für das Maß Wölbung sehr unterschiedliche Formen haben können [3], so dass sich konkrete Werte für die Wölbung einer Verteilung nur schwer interpretieren lassen.

Die Maßzahlen β_1 , β_2 bzw. γ_2 werden unter Verwendung der empirischen Momente höherer Ordnung geschätzt. Für die Realisierungen x_1, \dots, x_n einer Stichprobe X_1, \dots, X_n ergeben sich also

$$\hat{\beta}_1 = \frac{1}{n \cdot s^3} \sum_i^n (x_i - \bar{x})^3$$

$$\hat{\beta}_2 = \frac{1}{n \cdot s^4} \sum_i^n (x_i - \bar{x})^4$$

mit $\bar{x} = \sum_i^n x_i$ und $s^2 = \frac{1}{n} \sum_i^n (x_i - \bar{x})^2$. Die empirischen Momente sind empfindlich gegenüber Ausreißern in den Daten. Dies führt in Einzelfällen zu falschen Schlussfolgerungen bezüglich verschiedener Verteilungseigenschaften [49].

Aufgrund dieser problematischen Eigenschaften der klassischen Maßzahlen werden in dieser Arbeit wie bei Handl [33] alternative quantilbasierte Selektorstatistiken verwendet, um die Schiefe und die Länge der Schwänze einer Verteilung zu schätzen.

Sei $X_p = F^{-1}(p)$ das p -Quantil einer Zufallsvariablen X mit der Verteilungs-

funktion F . Ein Schätzer für die Schiefe wird über ein Verhältnis von Differenzen zwischen bestimmten Quantilen definiert:

$$S_q = \frac{X_{0.975} - X_{0.500}}{X_{0.500} - X_{0.025}}.$$

Diese Maßzahl kann nicht negativ werden und nimmt den Wert 1 für symmetrische Verteilungen an, da in diesem Fall das 97.5%-Quantil genauso weit vom Median entfernt ist wie das 2.5%-Quantil. Bei rechtsschiefen Verteilungen ist die Differenz im Zähler größer als diejenige im Nenner, also $S_q > 1$. Entsprechend gilt für linksschiefe Verteilungen $S_q < 1$. Unter Verwendung anderer Quantile lässt sich auch eine Maßzahl für die Länge der Verteilungsschwänze definieren:

$$T_q = \frac{X_{0.975} - X_{0.025}}{X_{0.875} - X_{0.125}}.$$

Für die Länge der Verteilungsschwänze gilt $T_q \geq 1$ und je größer der Wert, desto länger sind die Schwänze der Verteilung.

Die Schätzung dieser beiden Selektorstatistiken erfolgt unter Verwendung des klassischen Quantilschätzers [33], der einfach zu berechnen ist und auf einer Glättung der empirischen Verteilungsfunktion beruht. Das p -Quantil lässt sich basierend auf einer geordneten Stichprobe $x_{(1)}, \dots, x_{(n)}$ schätzen durch

$$x_p = \begin{cases} x_{(1)} & \text{für } p < 0.5/n \\ (1 - \lambda)x_{(j)} + \lambda x_{(j+1)} & \text{für } 0.5/n \leq p \leq 1 - 0.5/n \\ x_{(n)} & \text{für } p > 1 - 0.5/n \end{cases}$$

mit $j = \lfloor np + 0.5 \rfloor$ und $\lambda = np + 0.5 - j$. Dieser Quantilschätzer hat gegenüber anderen Schätzern eine hohe Effizienz [33].

Das von Hogg vorgeschlagene adaptive Verfahren verwendet jedoch Selektorstatistiken, die auf Differenzen asymmetrisch getrimmter Mittelwerte basieren. Die theoretische Selektorstatistik für die Schiefe ist definiert als

$$S_m = \frac{\mu_{0.95,1} - \mu_{0.25,0.75}}{\mu_{0.25,0.75} - \mu_{0,0.05}},$$

wobei $\mu_{a,b}$ den Mittelwert zwischen $F^{-1}(a)$ und $F^{-1}(b)$ bezeichnet. Die entsprechende Maßzahl für die Länge der Verteilungsschwänze lautet:

$$T_m = \frac{\mu_{0.95,1} - \mu_{0,0.05}}{\mu_{0.50,1} - \mu_{0,0.50}}.$$

Die Interpretation ist analog zu den quantilbasierten Selektorstatistiken.

Bei der Schätzung dieser Selektorstatistiken wird für $\mu_{a,b}$ der entsprechend getrimmte Mittelwert verwendet. D.h. es werden zunächst $a \cdot n$ der kleinsten sowie $(1 - b) \cdot n$ der größten Werte entfernt, bevor der Mittelwert bestimmt wird.

Diese Statistiken haben jedoch einige Nachteile [33]. Sie existieren nicht für alle Verteilungen, und die Interpretation der Mittelwerte ist zum Teil schwierig. Außerdem lässt sich die asymptotische Verteilung der Schätzer nicht einfach bestimmen.

3.2.2. Lineare Rangstatistiken

Unter Verwendung linearer Rangstatistiken lassen sich nichtparametrische Tests definieren, die für bestimmte Verteilungsformen optimal sind. Auch der Kruskal-Wallis-Test (siehe 2.2.1) und der Median-Test (siehe 2.2.1) lassen sich als lineare Rangstatistik angeben. Für die Berechnung dieser Teststatistik werden zunächst die Ränge in der kombinierten Stichprobe X_1, \dots, X_n vergeben, d.h. $X_{(1)}, \dots, X_{(n)}$ bezeichnet die geordnete Gesamtstichprobe. Die Indikatorvariable V_{ik} gibt die Stichprobenzugehörigkeit an:

$$V_{ik} = \begin{cases} 1 & \text{falls } X_{(k)} \text{ zum Genotyp } i \text{ gehört} \\ 0 & \text{sonst} \end{cases}$$

Für die Teststatistik werden von den Rängen $k = 1, \dots, n$ abhängige reelwertige Gewichte $a(k)$, $k = 1, \dots, n$ verwendet. Für jeden Genotyp i wird eine Statistik a_i definiert, die das durchschnittliche Gewicht für den i -ten Genotyp angibt:

$$a_i = \frac{1}{n_i} \sum_{k=1}^n a(k) V_{ik}.$$

Die lineare Rangstatistik ergibt sich dann als

$$L = \frac{(n-1) \sum_{i=0}^2 n_i (a_i - \bar{a})^2}{\sum_{k=1}^n (a(k) - \bar{a})^2},$$

wobei $\bar{a} = \frac{1}{n} \sum_{k=1}^n a(k)$ den empirischen Gesamt-Mittelwert der Gewichte bezeichnet. Diese Teststatistik vergleicht zwar ebenso wie die Statistik der Varianzanalyse quadrierte Abweichungen, jedoch wird im Nenner nicht die Variabilität innerhalb eines Genotyps betrachtet. Stattdessen gehen hier die Differenzen der einzelnen Gewichte zum Gesamt-Mittelwert ein, so dass der Wert des Nenners nicht von der Stichprobeneinteilung, sondern nur von der Gesamtfallzahl und den Gewichten abhängt. Unter der Nullhypothese gleicher Verteilungsformen und damit auch gleicher Mittelwerte bzw. Mediane ist L verteilungsfrei und asymptotisch χ^2 -verteilt mit $c - 1 = 2$ Freiheitsgraden [77].

Verschiedene lineare Rangtests, die jeweils besonders für bestimmte Verteilungsformen geeignet sind, können durch unterschiedliche Gewichte definiert werden [12]. In der Abbildung 3.1 sind die Gewichte für verschiedene Tests dargestellt, die in dieser Arbeit verwendet werden.

Kruskal-Wallis-Test

Die Teststatistik des Kruskal-Wallis-Test aus Abschnitt 2.2.1 lässt sich als lineare Rangstatistik darstellen, bei der das Gewicht dem jeweiligen Rang entspricht:

$$a_{KW}(k) = k$$

Dieser Test ist optimal, wenn die Verteilungen mittlere bis lange Verteilungsschwänze haben [12].

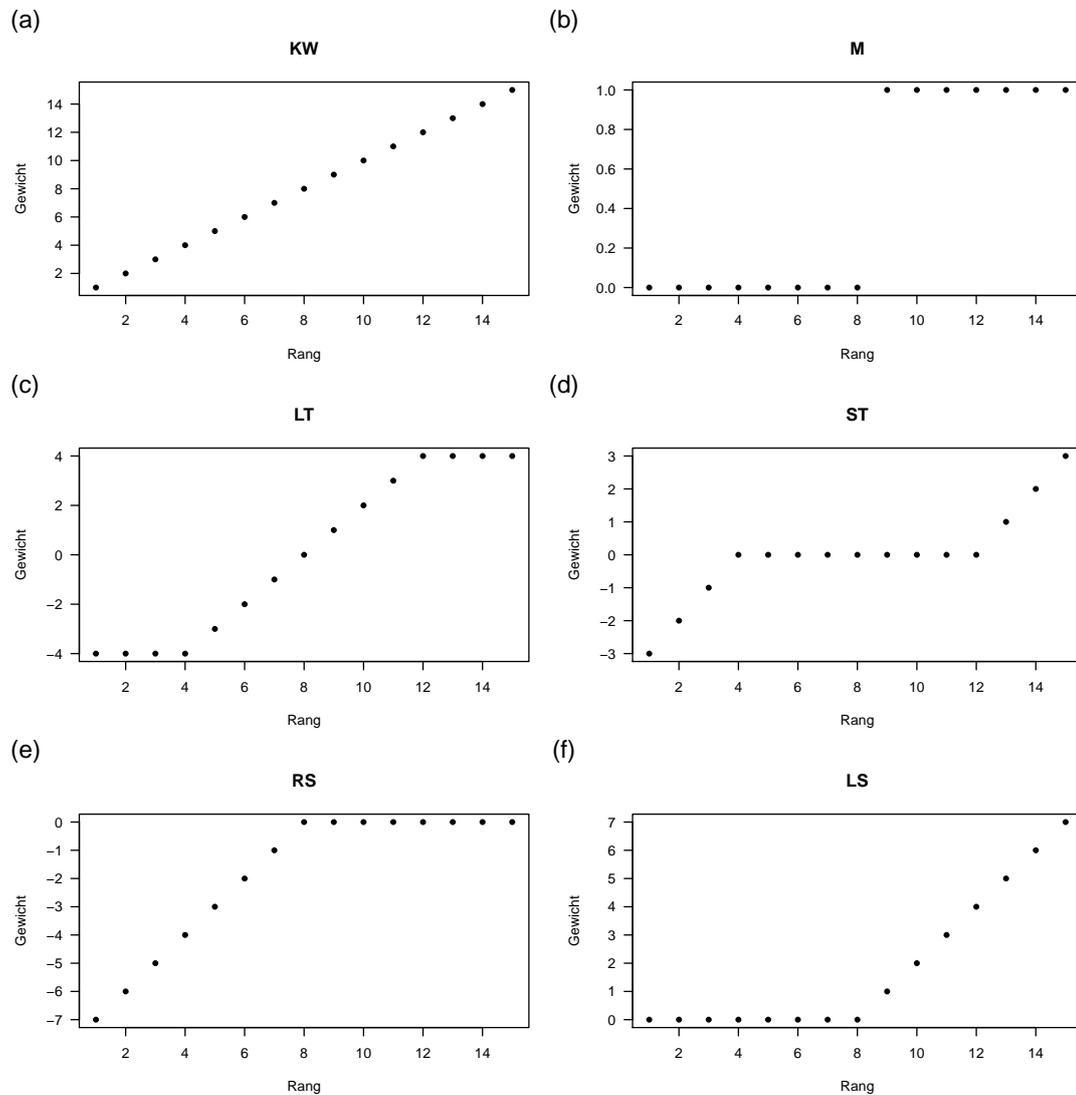


Abbildung 3.1: Gewichte für verschiedene lineare Rangtests: (a) Kruskal-Wallis-Test, (b) Median-Test, (c) Long-Tails-Test, (d) Short-Tails-Test, (e) Right-Skewness-Test und (f) Left-Skewness-Test. Die Wertebereiche auf den y-Achsen sind unterschiedlich.

Median-Test

Auch der Median-Test lässt sich als linearer Rangtest definieren, wobei hier als Gewicht nur die Information verwendet wird, ob der Rang kleiner oder größer als der mittlere Rang ist:

$$a_M(k) = \begin{cases} 1 & k > \frac{n+1}{2} \\ 0 & k \leq \frac{n+1}{2} \end{cases}$$

Er ist für Verteilungen mit sehr langen Schwänzen geeignet [12].

Test für lange Verteilungsschwänze

Speziell für die Situation langer Verteilungsschwänze wurde der Long-Tails-Test (LT-Test) entwickelt [14]. Hier gehen die Ränge entsprechend ihrer Abweichung vom mittleren Rang in die Teststatistik ein, wobei jedoch die Gewichte für die extremsten Ränge beschränkt sind:

$$a_{LT}(k) = \begin{cases} -(\lfloor \frac{n}{4} \rfloor + 1) & \text{falls } k < \lfloor \frac{n}{4} \rfloor + 1 \\ k - \frac{n+1}{2} & \text{falls } \lfloor \frac{n}{4} \rfloor + 1 \leq k \leq \lfloor \frac{3(n+1)}{4} \rfloor \\ \lfloor \frac{n}{4} \rfloor + 1 & \text{falls } k > \lfloor \frac{3(n+1)}{4} \rfloor \end{cases}$$

Test für kurze Verteilungsschwänze

Der Short-Tails-Test (ST-Test) wurde von Gastwirth [28] für kurze Verteilungsschwänze vorgeschlagen und wird deshalb auch als Gastwirth-Test bezeichnet. Hier werden nur die kleinsten und die größten 25% der Ränge in der Teststatistik verwendet:

$$a_{ST}(k) = \begin{cases} k - \frac{n+1}{4} & \text{falls } k \leq \frac{n+1}{4} \\ 0 & \text{falls } \frac{n+1}{4} < k < \frac{3(n+1)}{4} \\ k - \frac{3(n+1)}{4} & \text{falls } k \geq \frac{3(n+1)}{4} \end{cases}$$

Test für rechtsschiefe Verteilungen

Bei allen bisherigen Tests war die Gewichtung der Ränge symmetrisch. Anders sieht es bei rechts- oder linksschiefen Verteilungen aus. Beim Right-Skewness-Test (RS-Test), auch Hogg-Fisher-Randles-Test genannt, gehen nur die Ränge, die kleiner als der mittlere Rang sind, in die Teststatistik ein:

$$a_{RS}(k) = \begin{cases} k - \frac{n+1}{2} & \text{falls } k \leq \frac{n+1}{2} \\ 0 & \text{falls } k > \frac{n+1}{2} \end{cases}$$

Dieser Test ist somit für rechtsschiefe Verteilungen geeignet, da die kleinsten Werte in diesem Fall am informativsten sind, um einen Lageunterschied zu entdecken [41].

Test für linksschiefe Verteilungen

Der Left-Skewness-Test (LS-Test) verwendet entsprechend nur die Ränge, die größer als der mittlere Rang sind:

$$a_{LS}(k) = \begin{cases} 0 & \text{falls } k \leq \frac{n+1}{2} \\ k - \frac{n+1}{2} & \text{falls } k > \frac{n+1}{2} \end{cases}$$

3.2.3. Adaptive lineare Rangtests

Zwei adaptive Tests, die auf den quantilbasierten Selektorstatistiken S_q und T_q und den linearen Rangtests basieren, wurden von Büning [15] und Beier [7] vorgeschlagen. Beide verwenden für die Schätzung der Schiefe und der Länge der Verteilungsschwänze die kombinierte Stichprobe, in der die Ränge vergeben und die Selektorstatistiken geschätzt werden. Unterschiedlich sind jedoch die Entscheidungsgrenzen, die zur Auswahl der linearen Rangtests führen. In Abbildung 3.2 sind diese Grenzen für beide Methoden dargestellt (für die genauen Werte s. Tabellen A.1 und A.2 im Anhang). Auf der x-Achse ist die Selektorstatis-

Die Statistik S_q für die Schiefe angegeben, wobei eine logarithmierte Darstellung gewählt wurde, um rechts- und linksschiefe Verteilung vergleichbar darstellen zu können. Die y-Achse gibt die Statistik T_q für die Länge der Verteilungsschwänze an. Für symmetrische Verteilungen werden von kurzen über mittellange bis hin zu langen Verteilungsschwänzen in beiden Fällen der ST-Test, der Kruskal-Wallis-Test und der LT-Test verwendet. Auch bei rechts- und linksschiefen Verteilungen werden die geeigneten Tests, der RS-Test bzw. der LS-Test ausgewählt. Unterschiede gibt es bei schiefen Verteilungen mit langen Schwänzen. Während der Test von Büning (Abbildung 3.2 (a)) hier den LT-Test nutzt, wird bei Beier (Abbildung 3.2 (b)) einer der Tests für schiefe Verteilungen genutzt.

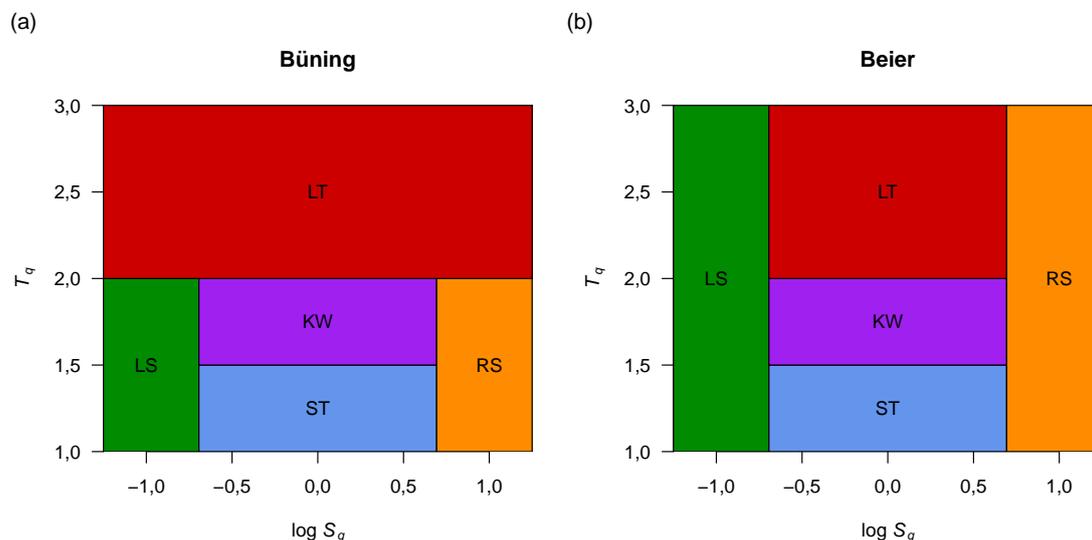


Abbildung 3.2: Adaptive lineare Rangtests von (a) Büning [15] und (b) Beier [7]. (ST: Short-Tails-Test, RS: Right-Skewness-Test, KW: Kruskal-Wallis-Test, LS: Left-Skewness-Test, LT: Long-Tails-Test)

Adaptive Verfahren, die auf diese Weise konstruiert werden, sind verteilungsfrei, d.h. sie halten das Signifikanzniveau ein. Dies wurde von Hogg [40] mit einem allgemein formulierten Satz gezeigt. Für eine Klasse F von Verteilungen gibt es k Tests, die auf den Teststatistiken T_1, \dots, T_k basieren und unter der Nullhypothese gleicher Verteilungsformen das Signifikanzniveau einhalten. Außerdem wird eine Selektorstatistik S definiert, die unter der Nullhypothese für alle Verteilungen in F unabhängig von den Teststatistiken T_1, \dots, T_k ist. Die Menge aller Werte für S kann in k paarweise disjunkte Teilmengen D_1, \dots, D_k aufgeteilt werden. Das adaptive Verfahren verwendet Test T_i , wenn der Wert für S in der Teilmenge D_i

liegt. Dieser zweistufige Test ist verteilungsfrei über F , d.h. er hält für alle Verteilungen in F das Signifikanzniveau ein.

Wenn als Verteilungsfamilie F alle stetigen Verteilungen definiert werden, können für T_1, \dots, T_k verschiedene lineare Rangtests verwendet werden, die für stetige Verteilungen unter der Nullhypothese gleicher Verteilungsformen das Signifikanzniveau einhalten. Die hier verwendeten Selektorstatistiken basieren auf den Rängen der kombinierten Stichprobe, d.h. sie sind Funktionen der geordneten Statistiken. Mithilfe eines Satzes von Basu (z.B. in [79]) lässt sich zeigen, dass sie unabhängig von den Statistiken der linearen Rangtests sind.

Ein Problem bei der Schätzung der Selektorstatistiken aus der kombinierten Stichprobe besteht darin, dass diese Schätzer nicht lage- und skaleninvariant sind [12] und somit unter der Alternativhypothese eventuell ein nicht optimaler Test ausgewählt wird. Alternativ können die Selektorstatistiken für jede Verteilung einzeln geschätzt werden. Auch Büning [13] verwendet stichprobenspezifische Selektorstatistiken S_q^i und T_q^i für die Genotypen $i = 0, 1, 2$, aus denen ein gewichtetes Mittel gebildet wird:

$$\bar{S}_q = \frac{1}{n} \sum_{i=0}^2 n_i \cdot S_q^i$$

$$\bar{T}_q = \frac{1}{n} \sum_{i=0}^2 n_i \cdot T_q^i$$

Diese Mittelwerte werden für die Auswahl eines geeigneten Tests verwendet. Nachteil ist, dass ein solcher adaptiver Tests nur noch asymptotisch verteilungsfrei ist [36]. In dieser Arbeit werden beide Möglichkeiten zur Schätzung der Selektorstatistiken untersucht.

3.3. Simulationen

Die hier beschriebene Simulationsstudie basiert auf dem gleichen Simulationsszenario wie Kapitel 2. Auch hier wird auf Unterschiede in der mittleren Expression der drei Genotypen getestet. Diese Simulationsstudie deckt zum einen

Verteilungen mit gleicher Form für die Expressionswerte der drei Genotypen ab. Zum anderen enthält sie zusätzlich Szenarien, in denen sich die Verteilungen entweder hinsichtlich ihrer Schiefe oder in der Länge ihrer Verteilungsschwänze unterscheiden. Um diese Parameter unabhängig voneinander variieren zu können, wird die Familie der g -und- k -Verteilungen verwendet. Diese Familie basiert auf den g -und- h -Verteilungen und wurde in verschiedenen Studien genutzt, um die Robustheit statistischer Verfahren bei Abweichungen von der Normalverteilung zu untersuchen [34, 48, 68].

3.3.1. Verteilungsfamilien

Die Familie der g -und- h -Verteilungen von Hoaglin [39] basiert auf einem Vorschlag von Tukey. Eine Übersicht der Eigenschaften findet sich in [65]. Die Quantile dieser Verteilungsfamilie ergeben sich aus einer Transformation der entsprechenden Quantile der Standardnormalverteilung:

$$a + b \left(\frac{1}{g} (\exp(gZ) - 1) \right) \exp(hZ^2/2), \quad (3.1)$$

wobei $Z = Z(p)$ das p -Quantil der Standardnormalverteilung angibt und a , $b > 0$, g und $h \geq 0$ Parameter für Lage, Variabilität, Schiefe und Länge der Verteilungsschwänze sind. Für $g = 0$ und $h = 0$ ergibt sich die Normalverteilung. Allgemein gilt für $g = 0$, $g < 0$ oder $g > 0$, dass die Verteilungen symmetrisch, linksschief oder rechtsschief sind. Bei Verteilungen mit $h = 0$ sind die Verteilungsschwänze so lang wie bei der Normalverteilung. Längere Verteilungsschwänze ergeben sich für $h > 0$. Verteilungen mit kürzeren Verteilungsschwänzen können jedoch nicht simuliert werden, da $h < 0$ nicht zu korrekten Verteilungen führt.

Eine Erweiterung ist die Familie der g -und- k -Verteilungen [78], deren Quantilfunktion definiert ist als

$$a + b \left(1 + c \left(\frac{1 - \exp(-gZ)}{1 + \exp(-gZ)} \right) \right) Z(1 + Z^2)^k. \quad (3.2)$$

Die Notation entspricht derjenigen der g -und- h -Verteilungen. Auch die Interpre-

tation ist analog. Jedoch hat diese Verteilungsfamilie zwei Vorteile. Zum einen sind die Parameter für die Schiefe und die Länge der Verteilungsschwänze bezüglich einer entsprechenden quantilbasierten Masszahl eher unabhängig. Zum anderen sind durch negative Werte für k auch Verteilungen mit kürzeren Verteilungsschwänzen als bei der Normalverteilung möglich. Hierbei sind jedoch symmetrische Verteilungen nur für $k > -\frac{1}{6}$ unimodal, während sie für $-\frac{1}{2} < k < -\frac{1}{6}$ bimodal sind. Im Fall von $g \neq 0$, sind die meisten Werte für k möglich [34].

Einen Überblick über die unterschiedlichen Dichten bei verschiedenen Werten für g und k gibt Abbildung 3.3. Im Teil (a) ist der Einfluss des Parameters g dargestellt, wenn die Verteilungsschwänze genauso lang sind wie bei der Normalverteilung ($k = 0$). Mit steigendem g ist die Verteilung stärker rechtsschief, bei negativem g würde sie entsprechend linksschiefer werden. In Teil (b) wird der Effekt von k bei symmetrischen Verteilungen ($g = 0$) gezeigt. Wenn k positiv ist, sind die Verteilungsschwänze länger und bei negativem k kürzer. Mit den g -und- k -Verteilungen lassen sich auch verschiedene bekannte Verteilungen approximieren [34].

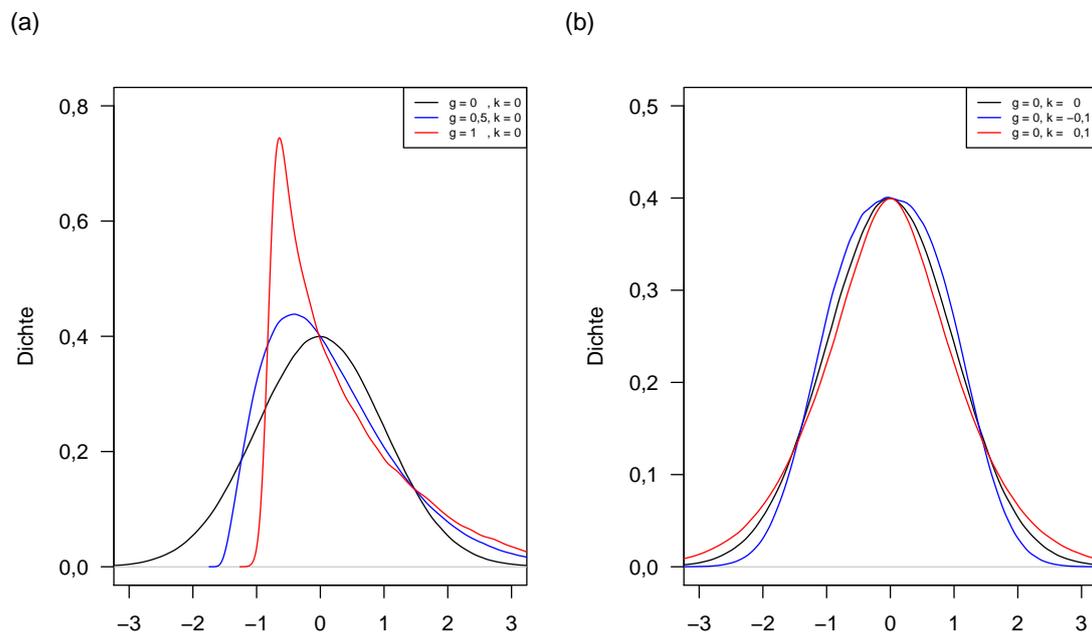


Abbildung 3.3: Dichten der g -und- k -Verteilungen für (a) unterschiedliche Schiefen (g) und (b) Längen der Verteilungsschwänze (k). Für die anderen Parameter gilt $a = 0$ und $b = 1$.

3.3.2. Simulierte Verteilungen

In dieser Simulationsstudie muss zwischen Situationen mit gleichen und unterschiedlichen Verteilungsformen unterschieden werden, die sich vor allem in der Anzahl der simulierten Szenarien und in der Darstellung der Ergebnisse unterscheiden.

Bei gleichen Verteilungsformen ist das jeweilige Szenario durch die zwei Parameter g und k festgelegt. Die Werte für g und k werden so gewählt, dass die Ebene aller Kombinationen möglichst gut abgedeckt wird. In Abbildung 3.4 sind die theoretischen Selektorstatistiken für die verwendeten Kombinationen von g und k dargestellt. Wie bei der Darstellung der adaptiven Verfahren werden die Schiefe und die Länge der Verteilungsschwänze auf der x- bzw. y-Achse abgetragen. Um die Werte besser einschätzen zu können, sind zusätzlich die theoretischen Selektorstatistiken für einige ausgewählte bekannte Verteilungen dargestellt. Die Laplace- und die verschiedenen t-Verteilungen haben z.T. deutlich längere Verteilungsschwänze als die Normalverteilung. Bei kleinen Freiheitsgraden der χ^2 -Verteilung ist die Rechtsschiefe und auch eine Verlängerung der Verteilungsschwänze erkennbar.

Bei den unterschiedlichen Verteilungsformen werden Schiefe und Länge der Verteilungsschwänze unabhängig voneinander betrachtet. Wie in Kapitel 2 werden auch hier sowohl ein Anstieg als auch ein Abfall der Schiefe bei den Genotypen 0, 1 und 2 simuliert. In Szenario I hat die Verteilung der Expression bei Genotyp 0 die kleinste Schiefe. Es werden zum einen Situationen simuliert, in denen die Schiefe von linksschief über symmetrisch nach rechtsschief variiert ($S_q = 0, 5, 1, 2$; $S_q = 0, 25, 1, 4$). Zum anderen gibt es Fälle, in denen die Verteilung der Expression bei Genotyp 0 symmetrisch, bei Genotyp 1 etwas rechtsschief und bei Genotyp 2 stärker rechtsschief ist ($S_q = 1, 2, 4$; $S_q = 1, 4, 8$). Beispiele für diese beiden Fälle für Szenario I sind in Abbildung 3.5 dargestellt. Szenario II deckt den umgekehrten Fall ab, bei dem Genotyp 0 die stärkste Schiefe aufweist. Auch hier gibt es wieder den Wechsel von rechtsschief über symmetrisch nach linksschief ($S_q = 2, 1, 0, 5$; $S_q = 4, 1, 0, 25$) sowie die unterschiedlich starken rechtsschiefen Verteilungen ($S_q = 4, 2, 1$; $S_q = 8, 4, 1$). Im Szenario III hingegen wird die Länge der Verteilungsschwänze variiert. In einigen Szenarien werden Verteilungen mit

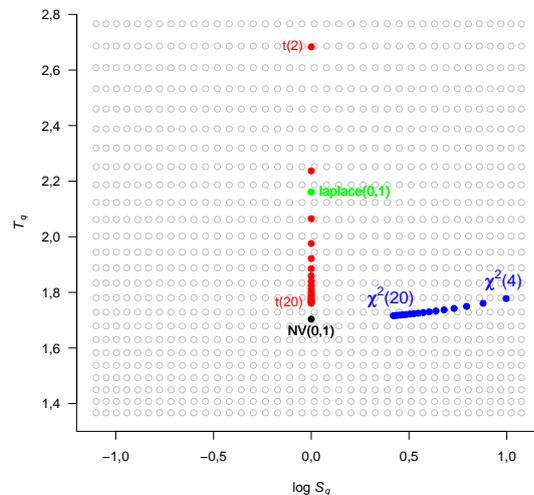


Abbildung 3.4: Theoretische Schiefe und Länge der Verteilungsschwänze für die simulierten gleichen Verteilungsformen (grau). Zum Vergleich sind die theoretischen Selektorstatisiken für die Normalverteilung (NV, schwarz), die Laplace-Verteilung (grün) sowie die t -Verteilung mit FG von 2 bis 20 (t , rot) und die χ^2 -Verteilung mit FG von 4 bis 20 (blau) dargestellt.

kurzen, mittleren und langen Schwänzen simuliert ($T_q = -0,17,0,0,22$). Außerdem werden Szenarien simuliert, in denen ausgehend von der Länge der Verteilungsschwänze der Normalverteilung zusätzlich zwei Verteilungen mit längeren Schwänzen verwendet werden ($T_q = 0,0,22,0,77$).

Da hier pro Szenario drei Parameter festgelegt werden, wird auf eine grafische Darstellung der Ergebnisse verzichtet und stattdessen werden wie in Abschnitt 2.4 Tabellen verwendet.

3.3.3. Parameter

In dieser Simulationsstudie wird ebenfalls ein SNP im Hardy-Weinberg-Gleichgewicht mit den Allelhäufigkeiten $p = 0,5$ und $p = 0,3$ simuliert. Zur besseren Unterscheidbarkeit der linearen Rangtests wird eine Gesamtfallzahl von $n = 200$ verwendet, so dass sich die absoluten Genotyp-Häufigkeiten $n_0 = n_2 = 50$ und $n_1 = 100$ für $p = 0,5$ sowie $n_0 = 98, n_1 = 84$ and $n_2 = 18$ für $p = 0,3$ ergeben.

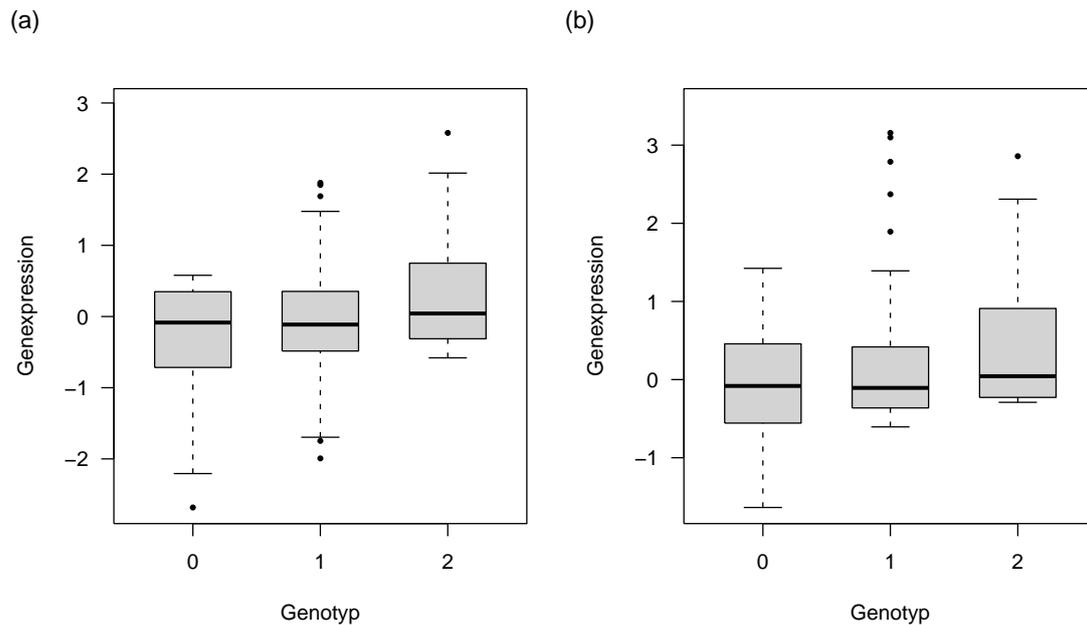


Abbildung 3.5: Beispiele für Simulationsszenario I der schiefen Verteilungen unter der Nullhypothese gleicher Mediane: (a) linksschief, symmetrisch, rechtsschief (b) symmetrisch, leicht rechtsschief, stark rechtsschief. Boxplots geben Median, Quartile, größten Nicht-Ausreisser und die Extrema an.

Hingegen werden, wie in der Simulationsstudie in Kapitel 2, unter der Alternativhypothese die Mediane 1, 1,25 und 1,5 für die Genotypen 0, 1 und 2 simuliert und für jedes Szenario 1000 Wiederholungen durchgeführt. Für die Beurteilung der empirischen Fehler 1. Art wird ebenfalls die Kriterien von Bradley verwendet.

3.4. Ergebnisse

Diese Simulationsstudie besteht aus zwei Teilen. Zunächst werden in den drei Gruppen die gleichen Verteilungsformen verwendet, um die Eigenschaften der verschiedenen linearen Rangtests und die Entscheidungsgrenzen der Methoden von Büning und Beier zu überprüfen. Basierend auf diesen Ergebnissen wird ein neuer adaptiver Test mit modifizierten Entscheidungsgrenzen vorgeschlagen. Anschließend werden die verschiedenen adaptiven Verfahren sowohl bei

gleichen als auch unterschiedlichen Verteilungsformen hinsichtlich ihrer empirischen Fehler 1. Art und empirischen Güte verglichen.

3.4.1. Entwicklung eines neuen adaptiven linearen Rangtests

Für die Abbildungen in diesem Abschnitt werden die empirischen Fehler und Gütwerte unter Verwendung der simulierten Szenarien aus Abbildung 3.4 geglättet.

Die linearen Rangtests halten in fast allen Szenarien das Signifikanzniveau sowohl für eine Allelhäufigkeit von 0,5 als auch von 0,3 ein (s. Abbildungen A.1 und A.2 im Anhang). Nur in einzelnen Fällen liegt der empirische Fehler 1. Art knapp über der Grenze von 7,5%.

In den Abbildungen 3.6 und 3.7 ist die empirische Güte über den gesamten Parameterraum dargestellt. Die Ergebnisse für die einzelnen Tests sind bei beiden Allelhäufigkeiten vergleichbar. Es wird jedoch deutlich, dass bei $p = 0,3$ die Güte insgesamt niedriger ist und dass beim Median-Test bei $p = 0,3$ die empirische Güte bei Rechtsschiefe höher ist als bei Linksschiefe. Die beiden Tests für rechts- bzw. linksschiefe Verteilungen (RS-Test und LS-Test) haben in der jeweiligen Schiefesituation die höchste Güte. Auch der ST-Test ist, wie im Abschnitt 3.2.2 beschrieben, am besten für Verteilungen mit kurzen Schwänzen geeignet. Der Kruskal-Wallis-Test zeigt ein ähnliches Verhalten, ist aber vor allem für Verteilungen mit normal langen Schwänzen mächtiger. Interessant ist, dass die Güteunterschiede beim LT-Test am geringsten sind und dieser noch nicht einmal bei langen Verteilungsschwänzen die höchste Güte hat. Das beim LT-Test erwartete Verhalten zeigt jedoch der Median-Test, wobei auch in dieser Simulation die insgesamt geringere Güte dieses Tests deutlich wird.

Abbildung 3.8 zeigt für jedes Szenario den jeweils optimalen Test, d.h. den Test mit der höchsten empirischen Güte. Obwohl die aus der Literatur bekannten Grenzen die Struktur einigermaßen abbilden, können sie jedoch noch optimiert werden. Zum einen fällt auf, dass der ST-Test für alle Szenarien mit kurzen Verteilungsschwänzen unabhängig von der Schiefe am mächtigsten ist. Zum anderen

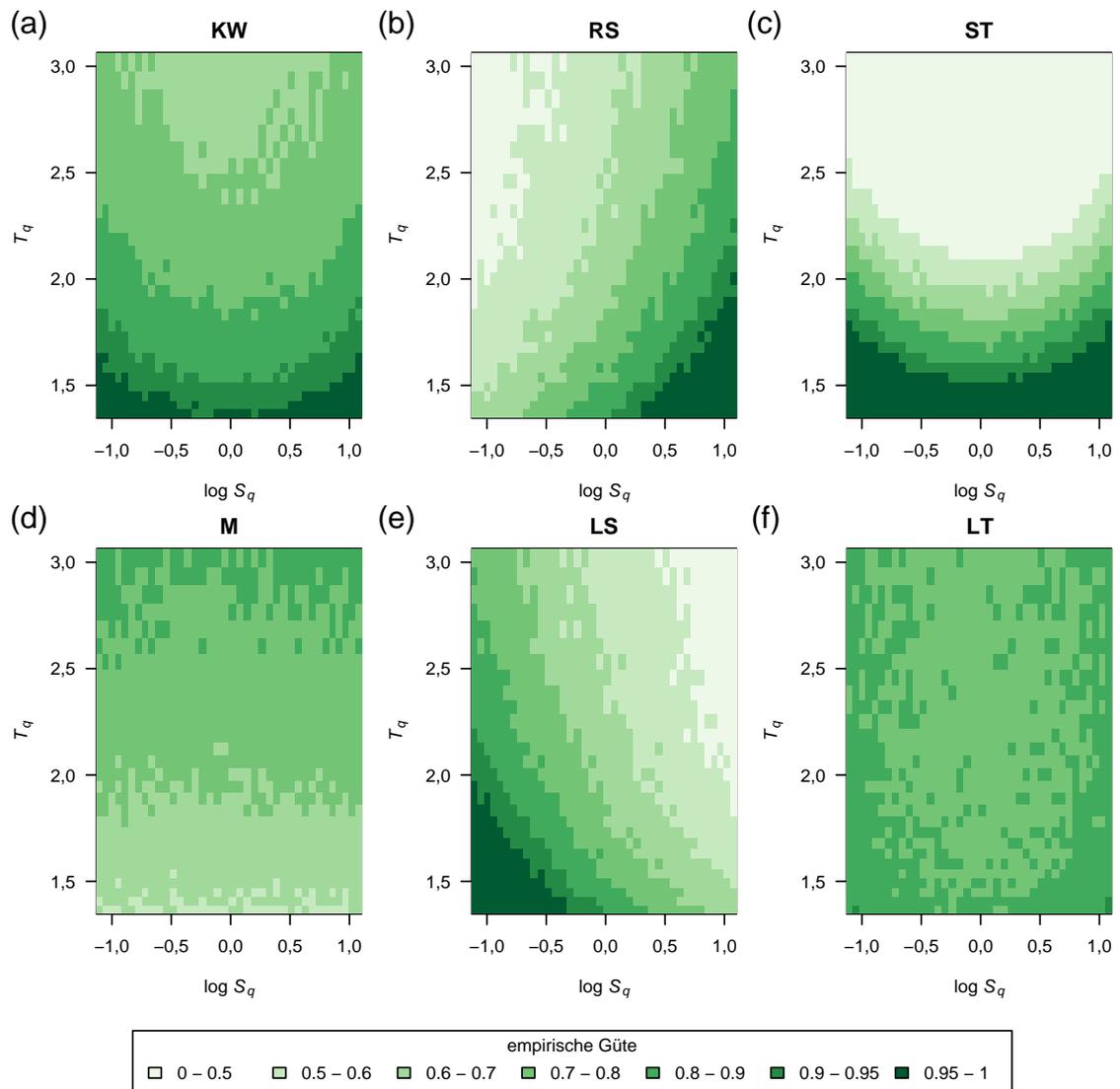


Abbildung 3.6: Empirische Güte der linearen Rangtests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,5: (a) Kruskal-Wallis-Test, (b) Right-Skewness-Test, (c) Short-Tails-Test, (d) Median-Test, (e) Left-Skewness-Test, (f) Long-Tails-Test

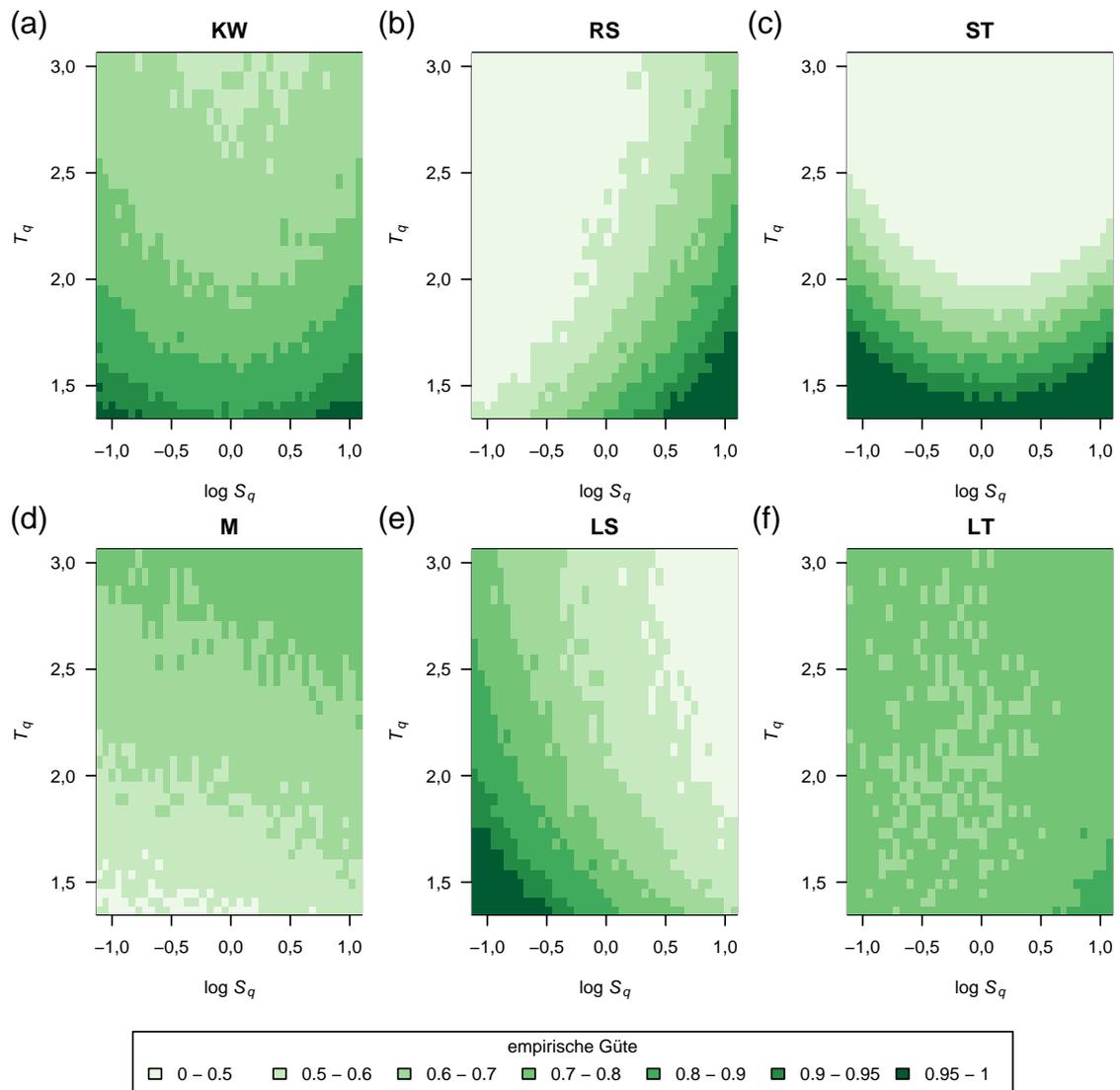


Abbildung 3.7: Empirische Güte der linearen Rangtests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,3: (a) Kruskal-Wallis-Test, (b) Right-Skewness-Test, (c) Short-Tails-Test, (d) Median-Test, (e) Left-Skewness-Test, (f) Long-Tails-Test

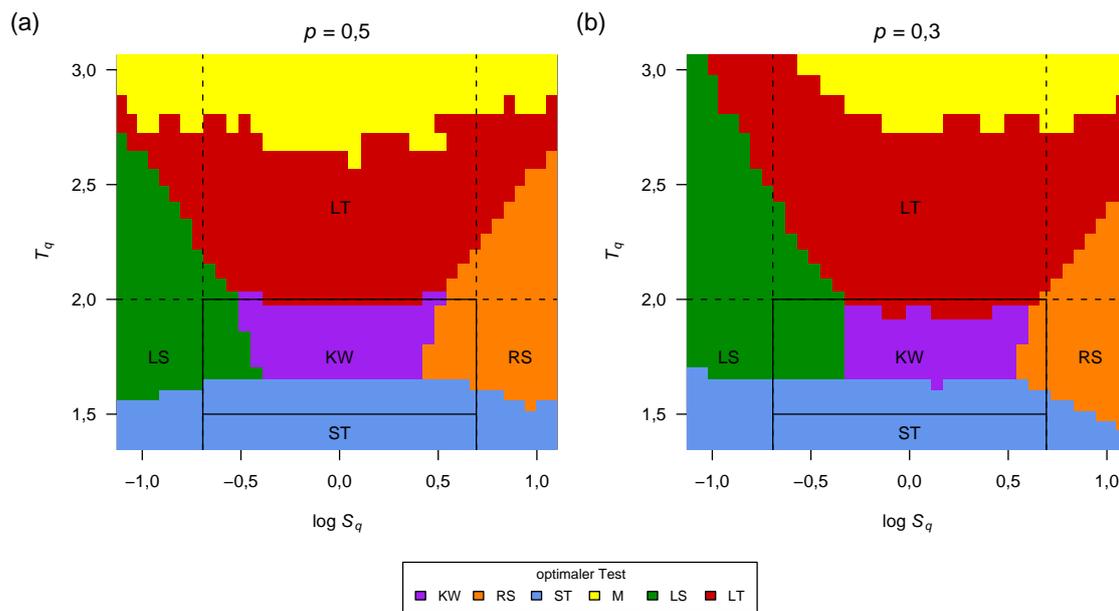


Abbildung 3.8: Optimale lineare Rangtests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz (a) 0,5 und (b) 0,3. Dargestellt sind zusätzlich die Klassifikations-schemata von Büning und Beier. (KW: Kruskal-Wallis-Test, LS: Left-Skewness-Test, LT: Long-Tails-Test, M: Median-Test, RS: Right-Skewness-Test, ST: Short-Tails-Test)

sollte für Situationen mit langen Verteilungsschwänzen die Entscheidung zwischen einem Test für Schiefe oder lange Verteilungsschwänzen sowohl von der Länge der Schwänze als auch von der Schiefe abhängig sein.

Basierend auf diesen Erkenntnissen wird ein neues Klassifikationsschema vorgeschlagen, das in Abbildung 3.9 dargestellt ist (s. auch Abbildung A.5 im Anhang für den optimalen linearen Rangtest mit dem neuen Schema). Hier wird für kurze Verteilungsschwänze unabhängig von der Schiefe der ST-Test verwendet. Außerdem wird bei Verteilungen mit normal langen Schwänzen schon bei relativ geringer Schiefe der LS-Test oder der RS-Test gewählt, während bei Verteilungen mit langen Schwänzen stärkere Schiefe nötig ist. Außerdem wird für Verteilungen mit sehr langen Schwänzen, die symmetrisch oder nur leicht schief sind, der Median-Test anstelle des Longtail Tests verwendet. Nicht geändert werden hingegen die Entscheidungsgrenzen zwischen dem Kruskal-Wallis-Test und dem LT-Test bzw. dem ST-Test. Während die erstgenannte Grenze durch die Simulationsergebnisse bestätigt wird, könnte letztere von 1,5 auf 1,7 verschoben werden. Dies hätte

jedoch den Nachteil, dass bei normalverteilten Daten die Selektorstatistik für die Länge der Verteilungsschwänze sehr nahe an dieser Grenze liegen würde.

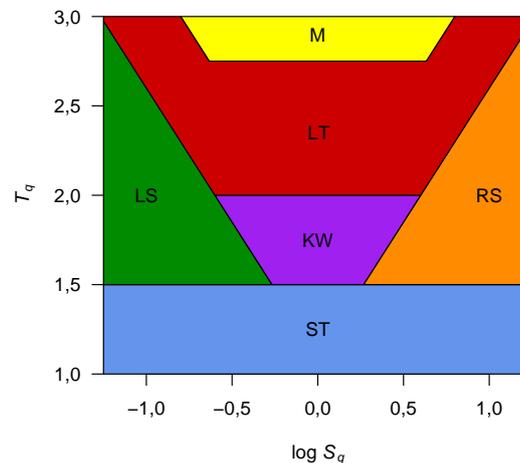


Abbildung 3.9: Neues adaptives Verfahren. (ST: Short-Tails-Test, RS: Right-Skewness-Test, KW: Kruskal-Wallis-Test, LS: Left-Skewness-Test, LT: Long-Tails-Test, M: Median-Test)

Wie im nächsten Abschnitt gezeigt wird, hält auch dieses neue adaptive Verfahren bei unterschiedlichen Verteilungsformen das Signifikanzniveau nicht ein. Eine Lösung für dieses Problem wäre ein Test, der in geeigneter Weise den Median-Test und den neuen adaptiven Test kombiniert. Zunächst müsste eine Entscheidung darüber getroffen werden, ob die Verteilungsformen in den drei Genotyp-Gruppen ähnlich genug sind, um den adaptiven Test anwenden zu können. Ist dies nicht der Fall, wird der Median-Test genutzt. Als Entscheidungskriterium könnte das Verhältnis der beiden extremen Selektorstatistiken für die Schiefe verwendet werden:

$$V_S = \frac{\max(S_q^i)}{\min(S_q^i)} \text{ für } i = 0, 1, 2.$$

Ein auf diese Weise konstruierter Test ist in Abbildung 3.10 dargestellt. Ist das Verhältnis der Selektorstatistiken größer als 5, wird der Median-Test verwendet und ansonsten ein geeigneter linearer Rangtest anhand der neuen Entscheidungsgrenzen ausgewählt. Für die Entscheidung zwischen ähnlichen oder unterschiedlichen Verteilungsformen wird der Wert 5 benutzt, da bei den Simulations-

szenarien mit extremen Schiefeunterschieden, d.h. einem theoretischen Schiefe-Verhältnis mehr als 8, 75 % der Replikate ein empirisches Schiefe-Verhältnis größer als 5 haben.

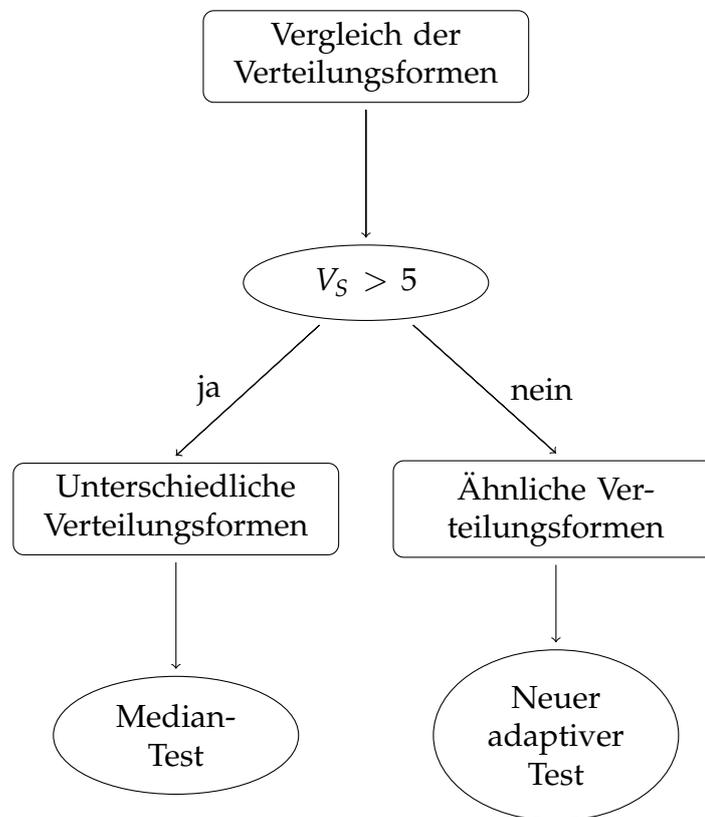


Abbildung 3.10: Kombination des Median-Tests und des neuen adaptiven Tests. Die Entscheidung für einen der beiden Tests basiert auf dem Verhältnis V_S der größten und kleinsten Genotyp-spezifischen Schiefe.

3.4.2. Vergleich der adaptiven Tests

Abschließend werden die verschiedenen adaptiven Tests hinsichtlich ihres empirischen Fehlerniveaus und ihrer empirischen Güte in Situationen mit gleichen und unterschiedlichen Verteilungen verglichen. Für jedes Verfahren (Büning, Beier, neues Schema und Kombination neues Schema mit Median-Test) werden die Selektorstatistiken sowohl basierend auf der kombinierten Stichprobe als auch separat geschätzt.

Gleiche Verteilungen

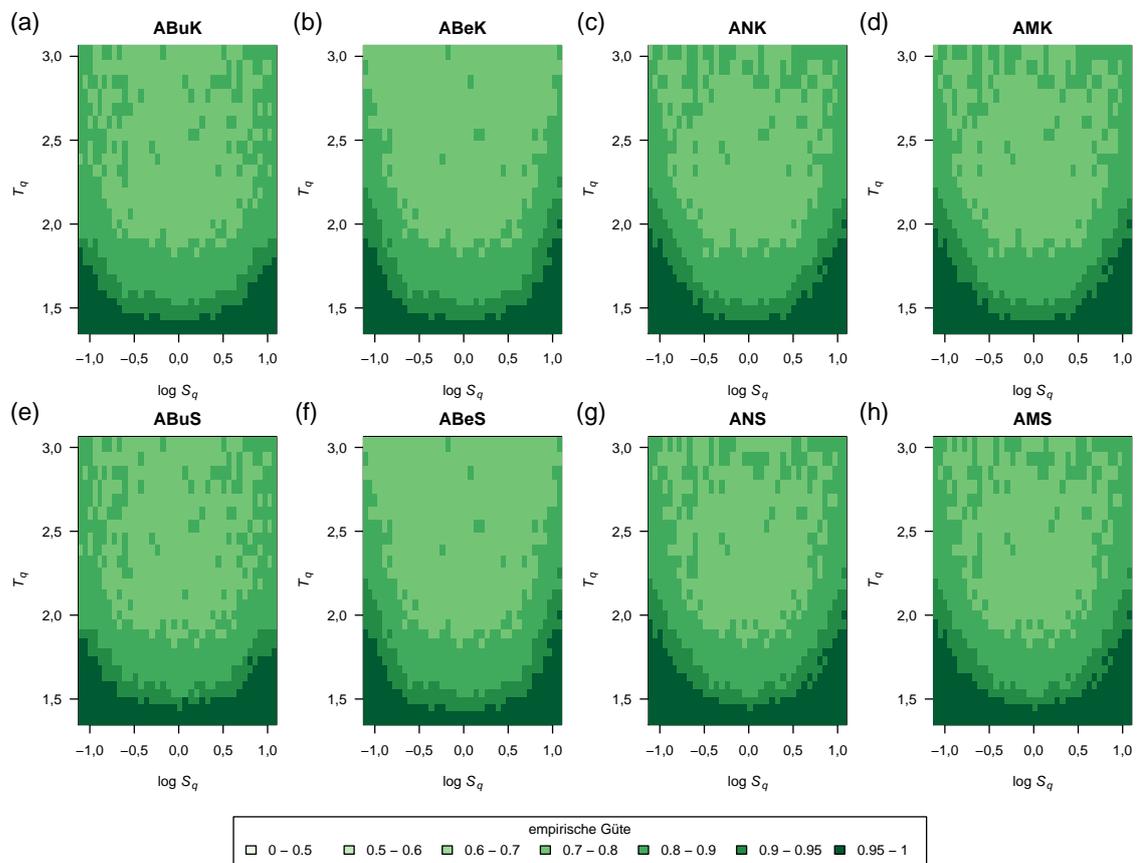


Abbildung 3.11: Empirische Güte der adaptiven Tests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,5: (a) Büning kombiniert, (b) Beier kombiniert, (c) Neuer Test kombiniert, (d) Kombiniertes Test kombiniert, (e) Büning separat, (f) Beier separat, (g) Neuer Test separat, (h) Kombiniertes Test separat

Alle adaptiven Tests sind bei gleichen Verteilungsformen robust, da die empirischen Fehler 1. Art bis auf wenige einzelne Ausnahmen innerhalb der Bradley-Kriterien liegen (s. Abbildungen A.3 und A.4 im Anhang).

Die empirische Güte, die in Abbildungen 3.11 und 3.12 dargestellt ist, unterscheidet sich für die beiden Methoden zur Schätzung der Selektorstatistiken, separat oder auf kombinierte Stichprobe, kaum. Bei allen Tests liegt die Güte bei Verteilungen mit kurzen und normal langen Verteilungsschwänzen zwischen 95 % und 100 %, während sie für symmetrische und leicht schiefe Verteilungen mit langen Schwänzen nur maximal 70 % beträgt.

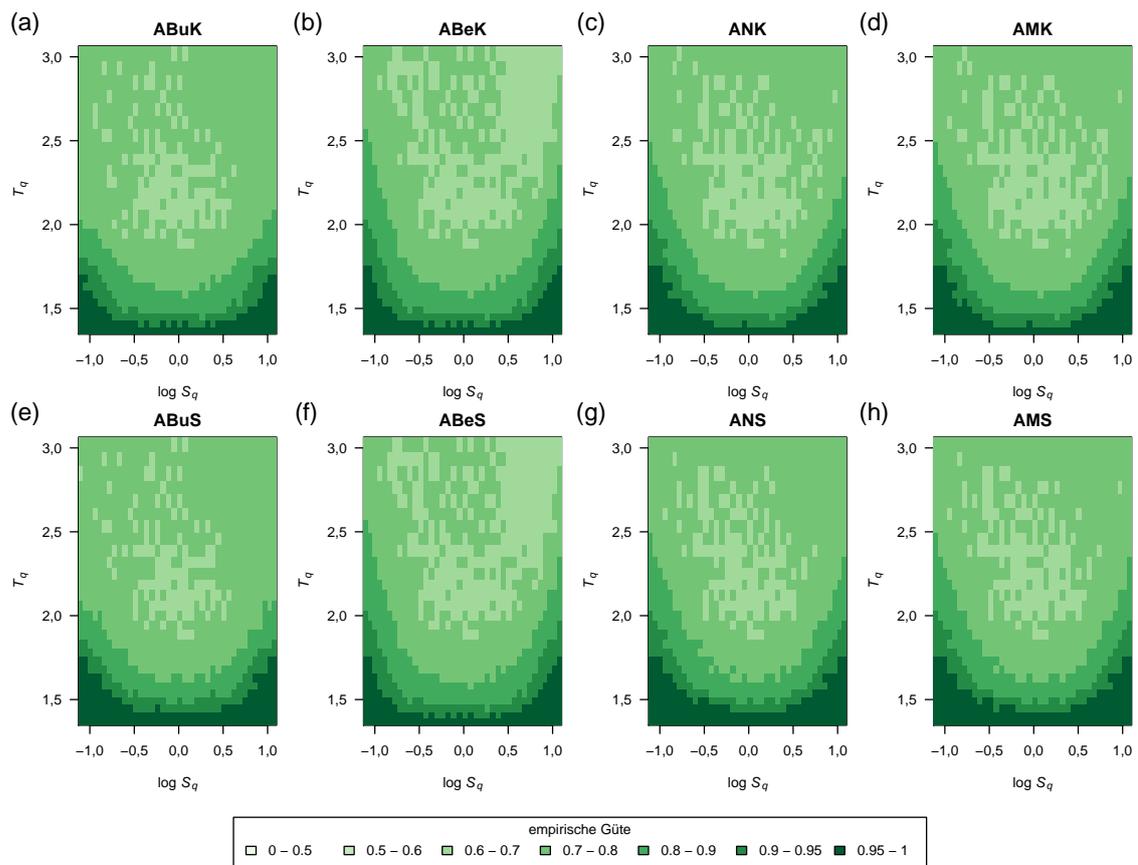


Abbildung 3.12: Empirische Güte der adaptiven Tests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,3: (a) Büning kombiniert, (b) Beier kombiniert, (c) Neuer Test kombiniert, (d) Kombiniertes Test kombiniert, (e) Büning separat, (f) Beier separat, (g) Neuer Test separat, (h) Kombiniertes Test separat

Die Unterschiede zwischen den adaptiven Verfahren sind klein, jedoch erkennbar. Der Test von Büning ist bei langen Verteilungsschwänzen besser als das Verfahren von Beier, dieses schneidet jedoch bei extrem schiefen Verteilungen besser ab. Das neue Schema hat in beiden Fällen eine hohe Güte. Bei einer Allelhäufigkeit von 0,5 sind die Ergebnisse des kombinierten Tests und des neuen adaptiven Tests nahezu identisch, bei einer geringeren Allelhäufigkeit von 0,3 hat der kombinierte Test jedoch bei Verteilungen mit kurzen Schwänzen z. T. eine geringe Güte.

Unterschiedliche Verteilungen

Tabelle 3.1 zeigt die empirischen Fehler 1. Art für die verschiedenen adaptiven Verfahren, sowie zum Vergleich die der Varianzanalyse, des Kruskal-Wallis-Tests und des Median-Tests. In den beiden Szenarien mit unterschiedlicher Schiefe (I und II) sind fast alle Methoden z.T. sogar extrem liberal, und nur der Median-Test ist robust. Der kombinierte Test hält in einigen der Szenarien mit großen Unterschieden in der Schiefe das Signifikanzniveau ein. Auch in den anderen Situationen sind die empirischen Fehler dieses Tests meistens kleiner als die empirischen Fehler der anderen adaptiven Tests. Beim Vergleich der beiden Schätzmethode für die Selektorstatistiken fällt auf, dass die empirischen Fehler höher sind, wenn die Selektorstatistiken separat geschätzt werden.

Im Gegensatz zum starken Einfluss der unterschiedlichen Schiefe halten alle Verfahren bei unterschiedlich langen Verteilungsschwänzen (Szenario III) das Signifikanzniveau ein.

In Tabelle 3.2 ist die empirische Güte unter der Alternativhypothese dargestellt. Für Szenarien I und II mit unterschiedlicher Schiefe ist kein fairer Vergleich der verschiedenen Methoden möglich, da nur der Median-Test das Signifikanzniveau einhält. In den vier Fällen, in denen auch der kombinierte adaptive Test robust ist, hat er eine vergleichbare empirische Güte wie der Median-Test. Bemerkenswert ist jedoch noch, dass der Median-Test in Szenario I die geringste empirische Güte hat, während er in den meisten Fällen von Szenario II am mächtigsten ist.

Bei Verteilungen mit unterschiedlich langen Verteilungsschwänzen (Szenario III) ist ein Vergleich aller Methoden möglich. Die verschiedenen adaptiven Tests haben eine sehr ähnliche Güte. Auch zwischen kombinierter und separater Schätzung der Selektorstatistiken ist kein Unterschied erkennbar.

Tabelle 3.1: Empirische Fehler 1. Art bei einem nominellen Signifikanzniveau $\alpha = 0.05$. Die verschiedenen Simulationsszenarien sind im Text beschrieben. Die Häufigkeit des selteneren Allels wird mit p bezeichnet, und S_q^i bzw. T_q^i geben die theoretische Schiefe und Länge der Verteilungsschwänze für Genotyp i an. Die statistischen Methoden sind die Varianzanalyse (VA), der Kruskal-Wallis-Test (KW), der Median-Test (M), die adaptiven Verfahren von Büning (ABu) und Beier (ABe) sowie der neu entwickelte Test einzeln (AN) und kombiniert (AM), wobei K und S die kombinierte bzw. separate Schätzung der Selektorstatistiken angeben.

p	S_q^0	S_q^1	S_q^2	T_q^0	T_q^1	T_q^2	VA	KW	M	ABuK	ABuS	ABeK	ABeS	ANK	ANS	AMK	AMS	
I	0,5	0,50	1,00	2,00	1,70	1,70	0,273	0,127	0,046	0,113	0,138	0,113	0,138	0,113	0,137	0,107	0,124	
		0,25	1,00	4,00			0,797	0,347	0,050	0,184	0,367	0,184	0,367	0,184	0,363	0,052	0,052	
		0,3	0,50	1,00	2,00			0,205	0,091	0,043	0,090	0,106	0,090	0,106	0,093	0,111	0,089	0,099
		0,25	1,00	4,00				0,670	0,277	0,045	0,136	0,317	0,143	0,317	0,146	0,325	0,060	0,067
		0,5	1,00	2,00	4,00			0,237	0,133	0,044	0,122	0,188	0,132	0,190	0,162	0,202	0,153	0,181
		0,3	1,00	2,00	4,00			0,546	0,366	0,047	0,375	0,608	0,518	0,621	0,538	0,641	0,168	0,180
II		1,00	2,00	4,00			0,178	0,093	0,053	0,091	0,118	0,091	0,118	0,087	0,117	0,085	0,107	
		1,00	4,00	8,00			0,556	0,290	0,056	0,193	0,423	0,215	0,433	0,237	0,458	0,168	0,222	
	0,5	2,00	1,00	0,50	1,70	1,70	0,244	0,092	0,037	0,087	0,109	0,087	0,109	0,087	0,110	0,090	0,105	
		4,00	1,00	0,25			0,803	0,331	0,046	0,176	0,362	0,176	0,362	0,176	0,346	0,048	0,049	
		0,3	2,00	1,00	0,50			0,214	0,093	0,047	0,086	0,107	0,086	0,107	0,089	0,125	0,090	0,118
		4,00	1,00	0,25				0,702	0,315	0,051	0,149	0,388	0,151	0,395	0,153	0,409	0,065	0,075
III		4,00	2,00	1,00			0,219	0,126	0,037	0,116	0,174	0,119	0,176	0,144	0,186	0,138	0,172	
		8,00	4,00	1,00			0,566	0,364	0,043	0,381	0,616	0,543	0,634	0,577	0,653	0,178	0,182	
	0,3	4,00	2,00	1,00			0,213	0,152	0,054	0,194	0,250	0,223	0,256	0,251	0,273	0,228	0,243	
		8,00	4,00	1,00			0,383	0,356	0,051	0,551	0,610	0,647	0,647	0,647	0,653	0,672	0,196	0,200
	0,5	1,00	1,00	1,00	1,50	1,70	2,00	0,058	0,062	0,044	0,062	0,062	0,062	0,061	0,062	0,061	0,062	0,061
					1,70	2,00	3,00	0,043	0,040	0,038	0,036	0,038	0,036	0,038	0,034	0,036	0,034	0,036
0,3				1,50	1,70	2,00	0,049	0,052	0,058	0,054	0,061	0,054	0,061	0,053	0,061	0,052	0,059	
				2,00	1,70	1,50	0,046	0,041	0,045	0,041	0,045	0,041	0,045	0,041	0,045	0,042	0,046	
				1,70	2,00	3,00	0,041	0,057	0,057	0,048	0,053	0,048	0,053	0,048	0,053	0,049	0,053	
				3,00	2,00	1,70	0,042	0,044	0,075	0,062	0,061	0,061	0,061	0,061	0,073	0,068	0,075	

Tabelle 3.2: Empirische Güte bei einem nominellen Signifikanzniveau $\alpha = 0.05$. Die verschiedenen Simulationsszenarien sind im Text beschrieben. Die Häufigkeit des selteneren Allels wird mit p bezeichnet, und S_i^j bzw. T_i^j geben die theoretische Schiefe und Länge der Verteilungsschwänze für Genotyp i an. Die statistischen Methoden sind die Varianzanalyse (VA), der Kruskal-Wallis-Test (KW), der Median-Test (M), die adaptiven Verfahren von Büning (ABu) und Beier (ABe) sowie der neu entwickelte Test einzeln (AN) und kombiniert (AM), wobei K und S die kombinierte bzw. separate Schätzung der Selektorstatistiken angeben.

	p	S_q^0	S_q^1	S_q^2	T_q^0	T_q^1	T_q^2	VA	KW	M	ABuK	ABuS	ABeK	ABeS	ANK	ANS	AMK	AMS	
I	0,5	0,50	1,00	2,00	1,70	1,70	1,70	(1,000)	(0,987)	0,711	(0,962)	(0,985)	(0,962)	(0,986)	(0,962)	(0,985)	(0,864)	(0,871)	
		0,25	1,00	4,00	(1,000)	(1,000)	(1,000)	(0,998)	(0,994)	0,846	(0,994)	(0,998)	(0,994)	(0,999)	(0,994)	(0,999)	0,846	(0,802)	0,846
	0,3	0,50	1,00	2,00	(0,995)	(0,954)	0,605	(0,914)	(0,952)	0,605	(0,914)	(0,952)	(0,914)	(0,952)	(0,913)	(0,954)	(0,802)	(0,821)	
		0,25	1,00	4,00	(1,000)	(0,998)	0,713	(0,966)	(0,996)	0,713	(0,967)	(0,996)	(0,967)	(0,996)	(0,963)	(0,995)	0,717	(0,869)	0,717
	0,5	1,00	2,00	4,00	(0,995)	(0,991)	0,740	(0,977)	(0,999)	0,740	(0,980)	(1,000)	(0,980)	(1,000)	(0,983)	(1,000)	(0,865)	(0,865)	
		1,00	4,00	8,00	(1,000)	(1,000)	0,861	(0,999)	(1,000)	0,861	(0,999)	(1,000)	(0,999)	(1,000)	(0,999)	(1,000)	(0,831)	(0,837)	
	0,3	1,00	2,00	4,00	(0,990)	(0,968)	0,676	(0,957)	(0,975)	0,676	(0,958)	(0,977)	(0,958)	(0,977)	(0,962)	(0,990)	(0,831)	(0,837)	
		1,00	4,00	8,00	(1,000)	(0,999)	0,958	(0,991)	(1,000)	0,958	(0,991)	(1,000)	(0,991)	(1,000)	(0,991)	(1,000)	(0,963)	(0,963)	
	II	0,5	2,00	1,00	0,50	(0,312)	(0,529)	0,608	(0,497)	(0,528)	0,608	(0,497)	(0,528)	(0,497)	(0,528)	(0,497)	(0,526)	(0,536)	(0,554)
			4,00	1,00	0,25	(0,063)	(0,214)	0,547	(0,283)	(0,223)	0,547	(0,283)	(0,223)	(0,283)	(0,220)	(0,285)	(0,213)	0,547	(0,474)
0,3		2,00	1,00	0,50	(0,283)	(0,452)	0,514	(0,415)	(0,442)	0,514	(0,415)	(0,442)	(0,415)	(0,441)	(0,400)	(0,431)	(0,453)	(0,474)	
		4,00	1,00	0,25	(0,071)	(0,202)	0,517	(0,268)	(0,208)	0,517	(0,259)	(0,208)	(0,259)	(0,191)	(0,253)	(0,186)	0,516	(0,578)	
0,5		4,00	2,00	1,00	(0,321)	(0,568)	0,629	(0,544)	(0,590)	0,629	(0,544)	(0,590)	(0,544)	(0,584)	(0,539)	(0,563)	(0,561)	(0,578)	
		8,00	4,00	1,00	(0,160)	(0,397)	0,563	(0,507)	(0,511)	0,563	(0,496)	(0,503)	(0,496)	(0,503)	(0,478)	(0,487)	(0,551)	(0,552)	
0,3		4,00	2,00	1,00	(0,249)	(0,483)	0,536	(0,481)	(0,488)	0,536	(0,473)	(0,473)	(0,473)	(0,475)	(0,434)	(0,459)	(0,496)	(0,509)	
		8,00	4,00	1,00	(0,136)	(0,425)	0,548	(0,601)	(0,602)	0,548	(0,600)	(0,602)	(0,600)	(0,600)	(0,564)	(0,575)	(0,564)	(0,564)	
III		0,5	1,00	1,00	1,00	1,50	1,70	2,00	0,824	0,831	0,648	0,830	0,827	0,830	0,827	0,828	0,825	0,828	0,825
			1,70	2,00	3,00	0,552	0,740	0,721	0,774	0,773	0,771	0,773	0,773	0,773	0,773	0,773	0,770	0,773	0,770
	0,3	1,50	1,70	2,00	0,756	0,783	0,549	0,776	0,786	0,776	0,786	0,776	0,786	0,786	0,770	0,785	0,765	0,780	
		2,00	1,70	1,50	0,740	0,759	0,593	0,758	0,760	0,758	0,760	0,758	0,760	0,759	0,759	0,762	0,759	0,762	
	1,70	2,00	3,00	0,523	0,681	0,600	0,683	0,681	0,683	0,681	0,683	0,681	0,681	0,683	0,680	0,682	0,679		
	3,00	2,00	1,70	0,442	0,643	0,673	0,703	0,701	0,699	0,701	0,699	0,701	0,700	0,701	0,695	0,701	0,696		

3.5. Diskussion

Bei gleichen Verteilungsformen in den Genotyp-Gruppen sind adaptive Tests eine hervorragende Alternative zu klassischen statistischen Methoden, wie der Varianzanalyse, dem Kruskal-Wallis-Test oder dem Median-Test. Durch die Schätzung der Selektorstatistiken wird ein für die konkrete Verteilungssituation optimaler Test ausgewählt und somit der Güteverlust des Median-Tests verhindert.

Wenn die Selektorstatistiken auf der kombinierten Stichprobe geschätzt werden und die Verteilungsformen gleich sind, kann bewiesen werden, dass die hier untersuchten adaptiven Tests das Signifikanzniveau einhalten. Bei einer separaten Schätzung gilt dies nur noch asymptotisch. Für die in dieser Studie simulierten hohen Fallzahlen lässt sich jedoch kein Unterschied in den empirischen Fehlern beim Vergleich der Schätzmethode feststellen. Eine offene Frage ist jedoch, ob dies auch für SNPs mit einer geringeren Allelhäufigkeit als 0,3 für das seltenere Allel gilt. Bei geringeren Allelhäufigkeiten könnte es außerdem von Vorteil sein, alternative Selektorstatistiken zu verwenden, die auf weniger extremen Quantilen basieren und mit einer geringeren Varianz geschätzt werden können.

Die in dieser Arbeit neu vorgeschlagenen Entscheidungsgrenzen kombinieren die Vorteile der beiden aus der Literatur bekannten Tests und führen somit insgesamt zu einer höheren Güte bei gleichen Verteilungsformen. Wenn die Verteilungen jedoch unterschiedlich sind, hält keines der adaptiven Verfahren das Signifikanzniveau ein. Eine mögliche Lösung ist der hier vorgestellte zweistufige Ansatz, der zunächst die Genotyp-spezifischen Selektorstatistiken für die Schiefe vergleicht, um zu entscheiden, ob die Verteilungsformen in den drei Genotyp-Gruppen unterschiedlich sind. Ist dies der Fall, wird der Median-Test verwendet, ansonsten das neue adaptive Entscheidungsschema. Dieser Test führt zwar in den meisten Szenarien zu den geringsten empirischen Fehlern, hält jedoch nur in einigen Situationen das Signifikanzniveau ein. Um den empirischen Fehler 1. Art auch in den anderen simulierten Szenarien zu reduzieren, könnte der Grenzwert für die Entscheidung kleiner als 5 gewählt werden. Dann würden die unterschiedlichen Verteilungen häufiger als solche erkannt werden. Allerdings würde es zu einem Güteverlust bei gleichen Verteilungsformen kommen, wenn diese fälschlicherweise als unterschiedlich klassifiziert werden würden.

4. Realdaten: Gutenberg-Herz-Studie

4.1. Zielsetzung

In den beiden vorhergehenden Kapiteln wurden die verschiedenen statistischen Methoden für eQTL-Studien mithilfe von Simulationen miteinander verglichen. Simulationen haben den Vorteil, dass die Wahrheit, in dieser Arbeit das Fehlen oder Vorhandensein von Assoziationen, bekannt ist. Jedoch können sie nicht alle in der Praxis auftretenden Fälle abdecken. Aus diesem Grund ist es wichtig, die Verfahren zusätzlich an realen Daten zu vergleichen.

In diesem Kapitel werden Genotyp- und Genexpressionsdaten aus der Gutenberg-Herz-Studie [103] verwendet, um zu untersuchen, welche eQTLs mit dem neuen adaptiven Verfahren im Vergleich zu klassischen Verfahren identifiziert werden. Um in solchen Studien die Anzahl der falsch-positiven Ergebnisse zu reduzieren, werden die p-Werte bzw. das Signifikanzniveau für die hohe Anzahl der durchgeführten Tests adjustiert [105]. Zusätzlich werden standardmäßig die Ergebnisse in mindestens einer unabhängigen Studie repliziert [71]. Um diese Vorgehensweise zu „simulieren“, werden die Probanden der Gutenberg-Herz-Studie künstlich in zwei Kohorten geteilt. Auf diese Weise können sowohl die Anzahl an identifizierten Assoziationen als auch der Anteil der replizierten Assoziationen bei den verschiedenen Methoden verglichen werden.

4.2. Daten

Die Gutenberg-Herz-Studie ist eine populationsbasierte, prospektive Kohortenstudie aus Mainz und dem Landkreis Mainz-Bingen, in der Risikofaktoren für kardiovaskuläre Erkrankungen identifiziert und untersucht werden. Ziel ist es, ein erhöhtes Risiko für diese Erkrankungen frühzeitig zu erkennen und mit passenden Therapien reagieren zu können. Die Teilnehmer der Studie werden umfassend klinisch, laborchemisch und epidemiologisch untersucht. Zusätzlich werden Genotyp- und Expressionsdaten erhoben, um genetische Komponenten für die Entstehung des frühzeitigen Herzinfarktes zu identifizieren. Die Genotypisierung erfolgt mit dem Affymetrix Genome-Wide Human SNP Array 6.0, der es ermöglicht, Genotypen von mehr als 900.000 SNPs zu bestimmen. In der Gutenberg-Herz-Studie werden für die Messung der Genexpression Monozyten aus frischen Blutproben verwendet, da diese einen großen Einfluss auf die Entstehung von Krankheiten des Immunsystems und Artherosklerose haben [96]. Die globale Expression von mehr als 48.000 Transkripten wird mit dem Illumina HT-12 v3 BeadChip ermittelt. Details zu den verwendeten Labormethoden sind in Zeller et al. [103] zu finden.

4.3. Methoden

Für die Einteilung der Probanden in zwei Kohorten wird das Datum der Erstuntersuchung verwendet, so dass die Fallzahlen in den beiden Gruppen ungefähr ein Verhältnis von 2:1 aufweisen. Um die Daten der beiden Kohorten möglichst unabhängig zu machen, müssten alle Vorverarbeitungsschritte sowohl bei den Genotyp- als auch bei den Expressionsdaten separat für jede Kohorte durchgeführt werden. Da jedoch die Bestimmung der Genotypen bei einer größeren Fallzahl genauer ist [43], wird für die Vorverarbeitung der Genotypen der komplette Datensatz verwendet. Im Gegensatz dazu hängen die Expressionswerte sehr stark davon ab, auf welchen Arrays die einzelnen Vorverarbeitungsschritte, vor allem Normalisierung, basieren [47]. Aus diesem Grund wird die Vorverarbeitung dieser Daten separat für beide Kohorten durchgeführt.

4.3.1. Vorverarbeitung der Genotypdaten

Nach der Genotypisierung mit dem Affymetrix Genome-Wide Human SNP Array 6.0 erhält man für jedes Array eine Datei mit Intensitätswerten für die beiden Allele jedes SNPs. Aus diesen Dateien werden zunächst die Genotypen mithilfe eines speziellen Algorithmus bestimmt (engl. genotype calling). In dieser Studie wird der Birdseed-Algorithmus von Affymetrix in der Version 2 verwendet [51].

Schlechte Qualität der DNA einzelner Personen und Fehler bei der Durchführung des Experiments können zu fehlerhaften Genotypdaten führen. Einige dieser Probleme können durch eine große Anzahl fehlender Genotypen oder einen hohen Anteil heterozygoter autosomaler SNPs erkannt werden. Aus diesem Grund werden Probanden entfernt, die mehr als 3 % fehlende Werte oder ein Heterozygoten-Verhältnis haben, das größer ist als das Dreifache der Standardabweichung des Verhältnisses für alle Arrays. Außerdem werden Personen entfernt, deren Genotypen sehr ähnlich sind. Hierfür wird paarweise der Anteil an gleichen Allelen (engl. identity-by-state) basierend auf 100.000 zufällig ausgewählten autosomalen SNPs geschätzt, und Personen mit Werten größer als 0,125 werden von weiteren Analysen ausgeschlossen.

Auf der SNP-Ebene gibt es ebenfalls eine Reihe von Standard-Kriterien, um problematische SNPs zu identifizieren [95, 105]. Zum einen sind seltene Varianten mit Array-basierten Technologien schwierig zu bestimmen, so dass SNPs mit einer relativen Häufigkeit des selteneren Allels (MAF, engl. minor allele frequency) kleiner als 1 % entfernt werden. Ein hoher Anteil an fehlenden Werten sowie eine Abweichung der relativen Genotyphäufigkeiten vom Hardy-Weinberg-Gleichgewicht deuten häufig auf Fehler bei der Bestimmung der Genotypen hin. Für statistische Analysen werden deswegen nur SNPs verwendet, die weniger als 1 % fehlende Genotypen aufweisen und deren p-Wert für einen Test auf Abweichung vom Hardy-Weinberg-Gleichgewicht kleiner als 10^{-4} ist.

Die beschriebenen Vorverarbeitungsschritte werden mit der Software PLINK durchgeführt [76].

4.3.2. Vorverarbeitung der Expressionsdaten

Expressionsarrays von Illumina verwenden die Bead-Technologie, bei der für jedes Transkript eine spezifische 50 bp lange Oligonukleotid-Sequenz an einer Glasperle (engl. bead) befestigt wird. Auf jedem Array werden ca. 30 Perlen des gleichen Typs zufällig verteilt [58]. Nach Hybridisierung und Scannen der Arrays erhält man für jedes Array eine Datei mit hintergrundkorrigierten Fluoreszenzintensitäten für jeden Perlentyp. Zunächst werden diese Werte zu einer mittleren Intensität pro Beadtyp zusammengefasst. Um die Abhängigkeit der Varianz vom Mittelwert für die verschiedenen Transkripte zu verringern, ist es üblich, die Intensitätswerte zu logarithmieren. In dieser Studie wird jedoch stattdessen eine speziell für Expressionsarrays entwickelte Varianz-stabilisierende Transformation verwendet [61]. Um die technische Variabilität zu minimieren, wird anschließend eine Quantilnormalisierung durchgeführt, die zu gleichen Verteilungen der Expressionswerte zwischen den Arrays führt [9].

Die Array-basierte Qualitätskontrolle besteht aus folgenden Schritten. Vor der Normalisierung werden zunächst Arrays mit einer schlechten Hybridisierungsqualität entfernt. Nach der Normalisierung werden außerdem technische Replikate, d.h. Arrays, die mit der gleichen RNA hybridisiert wurden, sowie Arrays mit Unterschieden im berichteten und expressionsbasierten Geschlecht ausgeschlossen. Für das expressionsbasierte Geschlecht wird die mittlere Expression von mehreren Genen verwendet, die auf dem Y-Chromosom außerhalb der pseudo-autosomalen Region liegen. Für Frauen erwartet man einen geringen Median nahe der Detektionsgrenze, während Männer einen höheren Median haben. Entsprechend werden Arrays mit einem unerwartet hohen oder niedrigen Median ausgeschlossen.

Transkripte werden basierend auf fehlender oder schlechter Annotation sowie geringen Expressionswerten gefiltert. Für weitere Analysen werden somit nur 25.682 Transkripte verwendet, deren SONDENSEQUENZ genau einmal im Referenzgenom vorhanden ist [22]. Geringe Expressionswerte deuten darauf hin, dass das jeweilige Transkript nicht zuverlässig gemessen werden kann und somit das Expressionsniveau dem Hintergrundrauschen entspricht. Für einen solchen Vergleich mit dem Hintergrund werden negative Kontrollproben verwendet, deren

Sequenzen nicht in der menschlichen DNA vorkommen. Der sogenannte Detektionsalgorithmus (engl. detection call algorithm) liefert einen p-Wert für jedes Transkript und jede Person [2]. In dieser Studie wird ein Transkript als detektierbar klassifiziert, wenn bei mindestens 10 % der Personen dieser p-Wert kleiner als 0,05 ist. Um Artefakte durch die Hybridisierungsmethode auszuschließen, werden für die eQTL-Analysen nur Transkripte verwendet, die keine SNPs in der Sondensequenz aufweisen [4].

Für die beschriebenen Vorverarbeitungsschritte werden die Bioconductor-Pakete `beadarray` [24] und `lumi` [22] genutzt.

4.3.3. Assoziationen zwischen SNPs und Transkripten

Für alle Transkripte, die in beiden Kohorten detektierbar sind, wird nach eQTLs mit SNPs in *cis* gesucht. Hierfür werden alle SNPs innerhalb des Gens sowie in einer Region von 500 kb um das Gen verwendet. Der Test auf Assoziation wird für jede dieser SNP-Transkript-Kombinationen mit der Varianzanalyse (nach Welch), dem Kruskal-Wallis-Test, dem Median-Test sowie dem kombinierten adaptiven Test durchgeführt, wobei bei letzterem die Schätzung der Selektorstatistiken auf der kombinierten Stichprobe durchgeführt wird. Um die Güte der Asymptotik zu verbessern, werden zum einen die SNPs neben der schon beschriebenen allgemeinen Qualitätskontrolle zusätzlich in beiden Kohorten separat nach der MAF gefiltert. Nur SNPs mit einer MAF größer als 5 % in jeder Kohorte werden für die eQTL-Analysen verwendet. Zum anderen werden bei SNPs mit einer geringen MAF die Homozygoten mit dem selteneren Allel und die heterozygoten Probanden zusammengefasst, wenn die erstgenannte Gruppe weniger als 15 Personen enthält.

In dieser Studie wird eine Assoziation als signifikant bezeichnet, wenn der p-Wert kleiner als 10^{-8} ist. Dies entspricht ungefähr der Bonferroni-Korrektur, da für jede Methode 2.374.883 Tests durchgeführt werden. Zusätzlich muss der Effekt, d.h. die maximale absolute Differenz zwischen den Medianen in den Genotyp-Gruppen, einen bestimmten Wert haben. Anstelle eines festen Grenzwerts für alle Transkripte wird er abhängig von der Gesamtvariabilität der Expressionswerte

des Transkripts festgelegt. Die Differenz muss mindestens genauso groß sein wie die Hälfte des Medians der absoluten Abweichungen (MAD, engl. median absolute deviation) des jeweiligen Transkripts. Für eine replizierte Assoziation müssen diese beiden Kriterien in beiden Kohorten separat erfüllt und zusätzlich noch das Vorzeichen der maximalen Differenz gleich sein.

4.4. Ergebnisse

Zunächst werden die beiden Kohorten anhand verschiedener Variablen miteinander verglichen, um die Ähnlichkeit zu zeigen. Anschließend werden die statistischen Methoden hinsichtlich der identifizierten und replizierten Assoziationen und Transkripten mit mindestens einer Assoziation verglichen. Für die SNPs in allen eQTLs, die mit dem adaptiven Test und nicht mit der Varianzanalyse gefunden werden, wird abschließend eine Abfrage im Katalog der GWA-Studien [37] durchgeführt, um einen Zusammenhang mit verschiedenen komplexen Erkrankungen und intermediären Phänotypen herzustellen.

4.4.1. Vergleich der beiden Kohorten

Tabelle 4.1: Vergleich der beiden Kohorten hinsichtlich Fallzahl, Alter, Anteil von Frauen, Anteil von Personen mit Myocard-Infarkt (MI) sowie Anzahl qualitätskontrollierter SNPs und Transkripte. Für das Alter sind Median und MAD angegeben, für die anderen Variablen neben der Anzahl ggf. der Anteil in Prozent.

Variable	Kohorte 1	Kohorte 2
Fallzahl	986	404
Alter	56 ± 13,34	52 ± 13,34
Weiblich	489 (50 %)	196 (49 %)
MI	19 (2 %)	11 (3 %)
SNPs	591.921	591.321
Transkripte	12.700	13.279

Tabelle 4.1 zeigt, dass die beiden Kohorten sehr ähnlich sind, obwohl Kohorte 2 nur ungefähr halb so groß ist wie Kohorte 1. Das mittlere Alter in beiden liegt bei Anfang bis Mitte 50, und die Hälfte der Probanden sind Frauen. Wie bei einer Kohortenstudie nicht anders zu erwarten, ist der Anteil an Personen, die zum Eintrittsdatum schon einen Herzinfarkt erlitten hatten, in beiden Kohorten gering. In beiden Kohorten stehen nach der Qualitätskontrolle 585.571 SNPs für weitere Analysen zur Verfügung. Von den 12.518 Transkripten, die in beiden Kohorten detektierbar sind, werden 2286 Transkripte ausgeschlossen, da sie mindestens einen SNP in ihrer SONDENSEQUENZ enthalten.

4.4.2. Vergleich der signifikanten und replizierbaren Assoziationen

Tabelle 4.2: Anzahl signifikanter und replizierter Assoziationen und Transkripte für die Varianzanalyse (VA), den Kruskal-Wallis-Test (KW), den Median-Test und den adaptiven kombinierten Test (AMK). Für die replizierten Ergebnisse ist zusätzlich der Anteil bezogen auf die signifikanten Ergebnisse in Kohorte 1 in Prozent angegeben.

	VA	KW	M	AMK
# sign. Assoziationen in Kohorte 1	27.057	28.861	20.093	27.828
# repl. Assoziationen in Kohorte 2	9720	10363	6524	9628
Anteil repl. Assoziationen in Kohorte 2	35,92 %	35,91 %	32,47 %	34,60 %
# sign. Transkripte in Kohorte 1	2165	2276	1779	2250
# repl. Transkripte in Kohorte 2	1068	1140	784	1106
Anteil repl. Transkripte in Kohorte 2	49,33 %	50,09 %	44,07 %	49,16 %

Die 10.232 Transkripte haben im Mittel 189 SNPs in *cis* (MAD: 92, Minimum: 5, Maximum: 1160). In Tabelle 4.2 sind die Anzahlen signifikanter und replizierter Assoziationen für die einzelnen Verfahren dargestellt. Außerdem werden die Methoden hinsichtlich der Transkripte mit mindestens einem eSNP (Expression-SNP), d.h. mit mindestens einer signifikanten Assoziation in *cis*, verglichen. Mit dem Median-Test werden am wenigsten Assoziationen und Transkripte identifiziert, während der Kruskal-Wallis-Test die höchste Anzahl aufweist. Mit dem adaptiven Test gibt es in Kohorte 1 mehr signifikante Assoziationen als bei der Varianzanalyse, jedoch wird in der zweiten Kohorte eine geringere Anzahl repli-

ziert. Im Gegensatz dazu liefert der adaptive Test eine höhere Anzahl an replizierten Transkripten. Insgesamt werden nur zwischen 32 % und 36 % der Assoziationen, von den meisten Methoden jedoch fast die Hälfte der Transkripte, auch in der zweiten Kohorte gefunden.

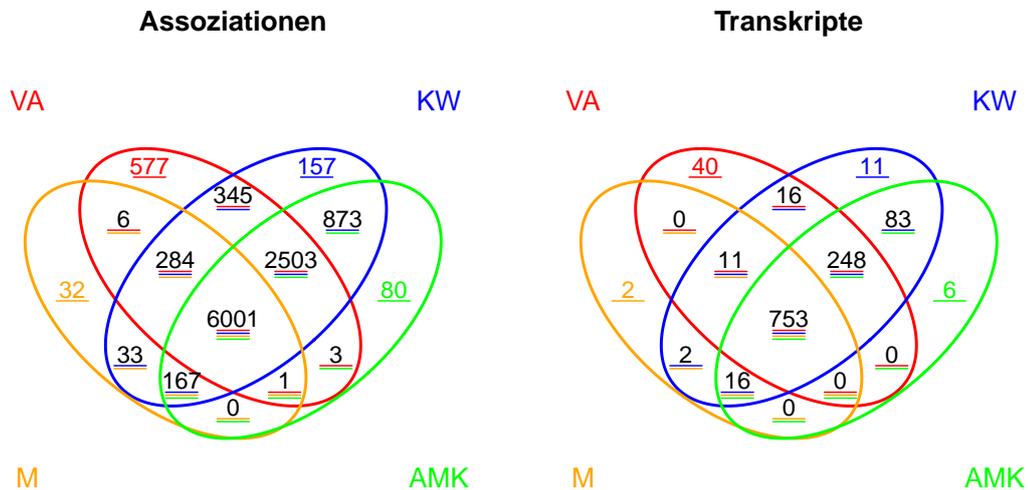


Abbildung 4.1: Venn-Diagramm für den Vergleich der reproduzierbaren Assoziationen (a) und Transkripte (b). Die statistischen Methoden sind die Varianzanalyse (VA), der Kruskal-Wallis-Test (KW), der Median-Test (M) und der adaptive kombinierte Test (AMK).

Das Venn-Diagramm in Abbildung 4.1 zeigt Unterschiede und Übereinstimmungen bei den identifizierten Assoziationen und Transkripten. Die Ergebnisse für Assoziationen und diejenigen für Transkripte mit mindestens einem eSNP sind ähnlich. Von den insgesamt 11.062 Assoziationen bzw. 1188 Transkripten werden 54 % bzw. 63 % von allen vier statistischen Methoden identifiziert. Die Varianzanalyse hat die höchste Anzahl an Methoden-spezifischen Signifikanzen, gefolgt vom Kruskal-Wallis-Test, dem adaptiven Test und dem Median-Test. Jedoch sind auch bei diesen Assoziationen die p-Werte der anderen Methoden in den meisten Fällen nur geringfügig kleiner als das festgelegte Signifikanzniveau von 10^{-8} . Wenn dieser Grenzwert auf 10^{-5} reduziert wird, werden 97 % bzw. 99 % der Assoziationen, die nur mit der Varianzanalyse bzw. dem Kruskal-Wallis-Test gefunden werden, von mindestens einem der anderen Verfahren identifiziert. Beim Median-Test und beim adaptiven Test beträgt dieser Anteil sogar 100 %.

4.4.3. Zusammenhang mit Ergebnissen aus GWA-Studien

Von den 1120 Assoziationen, die nicht mit der Varianzanalyse, sondern nur mit dem adaptiven Test gefunden werden, gibt es für sieben SNPs in fünf Regionen signifikante Assoziationen mit verschiedenen Krankheiten und intermediären Phänotypen, die in Tabelle 4.3 dargestellt sind. Zusätzlich enthält die Tabelle Informationen zu den identifizierten eQTL Assoziationen. Auch hier wird deutlich, dass sich die p-Werte der statistischen Methoden nicht sehr stark unterscheiden. In fast allen Fällen sind die p-Werte der Varianzanalyse in Kohorte 2 nur geringfügig größer als das Signifikanzniveau von 10^{-8} . Eine Ausnahme ist die Assoziation des SNPs rs867186 mit dem Gen *EIF6*, da diese auch nicht mit dem Kruskal-Wallis-Test und dem Median-Test, sondern nur mit dem adaptiven Test gefunden wird. Abbildung 4.2 zeigt die p-Werte der verschiedenen Methoden für dieses Gen. Alle Verfahren finden bei Kohorte 1 eQTL-Assoziationen in einer großen Region zwischen ungefähr 33,6 und 33,9 Mb. Jedoch sind diese Assoziationen bis auf einen SNP beim Kruskal-Wallis-Test nur beim adaptiven Test auch in der Kohorte 2 signifikant, wobei letzterer den LS-Test auswählt.

Für zwei der aufgeführten SNPs ist bekannt, dass sie die Expression des entsprechenden Gens beeinflussen. Zusätzlich stimmt die Richtung des Effekts der Genotypen auf die Expression in der jeweiligen Studie mit dem in der Gutenberg-Herz-Studie beobachteten Effekt überein. Der SNP rs599839 auf Chromosom 1 ist mit koronarer Herzkrankheit und verschiedenen intermediären kardiovaskulären Phänotypen assoziiert. Zusätzlich ist er in Lebergewebe von Kauasiern mit der Expression von *PSRC1* korreliert, wobei das Risikoallel zu einer geringeren Expression führt [82]. Für den mit Rheumatoider Arthritis assoziierten SNP rs3093024 auf Chromosom 6 konnte in Epstein-Barr Virus transfizierten Zelllinien von japanischen Individuen eine Assoziation mit der *CCR6*-Expression gezeigt werden, wobei, genau wie in der Gutenberg-Herz-Studie, die Expression mit der Anzahl an selteneren Allelen ansteigt [50].

Im Gegensatz dazu werden für den SNP rs8042680 im *PRC1*-Gen, der mit Typ-2-Diabetes mellitus assoziiert ist, unterschiedliche eQTL-Assoziationen berichtet. Während dieser SNP in der Gutenberg-Herz-Studie eine Assoziation mit der Expression des stromaufwärts gelegenen Gens *RCCD1* zeigt, hat er in isländischen

Blutproben einen Einfluss auf die Expression des stromabwärts gelegenen Gens *VPS33B* [93]. In der Gutenberg-Herz-Studie wird jedoch von keiner der Methoden eine Assoziation zwischen rs8042680 und dem *VPS33B*-Gen identifiziert.

4.5. Diskussion

Die Gutenberg-Herz-Studie mit einer hohen Anzahl an Probanden, von denen sowohl Genotyp- als auch Expressionsdaten vorhanden sind, ermöglicht einen Vergleich der verschiedenen in dieser Arbeit untersuchten statistischen Methoden. Die meisten Assoziationen und Transkripte mit mindestens einer Assoziation *in cis* werden vom Kruskal-Wallis-Test identifiziert. Methoden-spezifische Assoziationen haben jedoch bei den anderen Methoden in den meisten Fällen p-Werte, die nur knapp unter der gewählten Grenze von 10^{-8} liegen. Dies deutet darauf hin, dass es sich wahrscheinlich nicht um falsch-positive Ergebnisse handelt, sondern dass eher die unterschiedliche Güte der Verfahren der Grund ist. Jedoch hätte man in diesem Fall erwartet, dass der adaptive Test eine höhere Güte als der Kruskal-Wallis-Test hat und daher mehr Assoziationen identifiziert hätte. Verschiedene Gründe für das umgekehrte Ergebnis sind denkbar. Ein Grund könnte sein, dass der adaptive Test basierend auf einem Signifikanzniveau von 5% entwickelt, für die Analyse der realen Daten jedoch eine wesentlich kleinere Grenze von 10^{-8} verwendet wurde. Das Güteverhalten von verschiedenen Tests kann sich jedoch je nach Wahl des Signifikanzniveaus ändern, so dass die Ergebnisse aus der Simulation nicht vollständig übertragbar sind. Ein anderer möglicher Grund ist ein großer Standardfehler bei der Schätzung der Selektorstatistiken, der dazu führt, dass ein suboptimaler Test gewählt wird. Um das Verhalten des neuen adaptiven Tests besser zu verstehen, sind also weitere Simulationsstudien nötig, die die Güte bei verschiedenen, kleineren Signifikanzniveaus sowie die Eigenschaften der Schätzer für die Selektorstatistiken weiter untersuchen.

In vielen GWA-Studien wird häufig für die signifikantesten Loci nach eQTL-Assoziationen *in cis* gesucht, um mögliche molekulare Mechanismen und Kandidatengene zu identifizieren. In dieser Analyse wird jedoch ausgehend von den Assoziationen, die nur mit dem neuen adaptiven Test und nicht mit der Varianz-

analyse gefunden werden, für die jeweiligen SNPs nach Ergebnissen im Katalog der GWA-Studien gesucht. Neben zwei GWA-SNPs, für die die entsprechenden eQTL-Assoziationen schon in der Literatur berichtet wurden, liefert der adaptive Test neue Assoziationen mit bisher nicht in diesem Zusammenhang betrachteten Transkripten. Jedoch stellen die hier untersuchten Assoziationen nur einen kleinen Teil der in diesen Ergebnissen enthaltenen Informationen dar. Weniger strenge Kriterien für eine Signifikanz, z.B. durch alleinige Verwendung einer Grenze für den p-Wert, und vor allem eine Analyse unter Verwendung des kompletten Datensatzes führen sicherlich zu einer größeren Anzahl an identifizierten Assoziationen und Transkripten. Außerdem beschränkte sich die hier durchgeführte Analyse nur auf *cis*-Assoziationen innerhalb von 500 Mb um das jeweilige Gen, so dass keine Assoziationen in *trans* identifiziert werden konnten.

Tabelle 4.3: Signifikante Assoziationen aus GWA-Studien für eSNPs, die mit dem adaptiven Test und nicht mit der Varianzanalyse identifiziert werden. Die Tabelle enthält den SNP mit Chromosom (Chr) und Position sowie den assoziierten Phänotyp. Zusätzlich sind verschiedene Informationen zu der eQTL-Assoziation angegeben. Neben dem Gensymbol gibt der Typ an, ob sich der jeweilige SNP im Gen oder in einer Region von 100 kb bzw. 500 kb um das Gen befindet. Außerdem sind für beide Kohorten die p-Werte für die Varianzanalyse (VA), den Kruskal-Wallis-Test (KW), den Median-Test (M) sowie den adaptiven Test (AMK) angegeben.

SNP	Chr	Position	Phänotyp	Gen	Typ	Kohorte	VA	KW	M	AMK
rs629301	1	109818306	Blutfette [91]	PSRC1	cis 100kb	1	$8,33 \cdot 10^{-27}$	$3,37 \cdot 10^{-34}$	$1,91 \cdot 10^{-25}$	$3,37 \cdot 10^{-34}$
rs599839	1	109822166	Koronare Herzkrankheit [80], LDL-Cholesterin [81, 94, 101], Phospholipase A2[88]	PSRC1	cis 100kb	2	$2,74 \cdot 10^{-8}$	$8,90 \cdot 10^{-13}$	$1,58 \cdot 10^{-10}$	$8,90 \cdot 10^{-13}$
rs3093024	6	167532793	Rheumatoide Arthritis [50]	CCR6	im Gen	1	$1,21 \cdot 10^{-26}$	$1,61 \cdot 10^{-33}$	$3,27 \cdot 10^{-25}$	$1,61 \cdot 10^{-33}$
rs3093023	6	167534290	Rheumatoide Arthritis [86]	CCR6	im Gen	2	$1,96 \cdot 10^{-8}$	$7,07 \cdot 10^{-12}$	$2,22 \cdot 10^{-9}$	$7,07 \cdot 10^{-12}$
rs8042680	15	91521337	Typ-2-Diabetes mellitus [93]	RCCD1	cis 100kb	1	$1,48 \cdot 10^{-42}$	$2,34 \cdot 10^{-39}$	$6,10 \cdot 10^{-28}$	$3,25 \cdot 10^{-29}$
rs3733829	19	41310571	Rauchverhalten [67]	EGLN2	im Gen	2	$7,98 \cdot 10^{-8}$	$1,59 \cdot 10^{-11}$	$6,43 \cdot 10^{-9}$	$2,55 \cdot 10^{-10}$
rs867186	20	33764554	Gerinnungsfaktoren [85]	EIF6	cis 500kb	1	$1,26 \cdot 10^{-37}$	$9,40 \cdot 10^{-37}$	$3,20 \cdot 10^{-26}$	$8,74 \cdot 10^{-28}$
						2	$4,86 \cdot 10^{-8}$	$6,17 \cdot 10^{-12}$	$1,28 \cdot 10^{-9}$	$9,43 \cdot 10^{-11}$
						1	$5,08 \cdot 10^{-17}$	$2,93 \cdot 10^{-16}$	$1,12 \cdot 10^{-12}$	$2,93 \cdot 10^{-16}$
						2	$2,28 \cdot 10^{-8}$	$4,00 \cdot 10^{-9}$	$4,52 \cdot 10^{-5}$	$4,00 \cdot 10^{-9}$
						1	$1,42 \cdot 10^{-18}$	$6,62 \cdot 10^{-27}$	$1,00 \cdot 10^{-22}$	$6,62 \cdot 10^{-27}$
						2	$5,76 \cdot 10^{-8}$	$5,04 \cdot 10^{-11}$	$1,60 \cdot 10^{-10}$	$2,08 \cdot 10^{-11}$
						1	$4,09 \cdot 10^{-16}$	$2,79 \cdot 10^{-18}$	$1,44 \cdot 10^{-11}$	$2,79 \cdot 10^{-18}$
						2	$1,57 \cdot 10^{-5}$	$2,16 \cdot 10^{-7}$	$5,07 \cdot 10^{-5}$	$9,50 \cdot 10^{-9}$

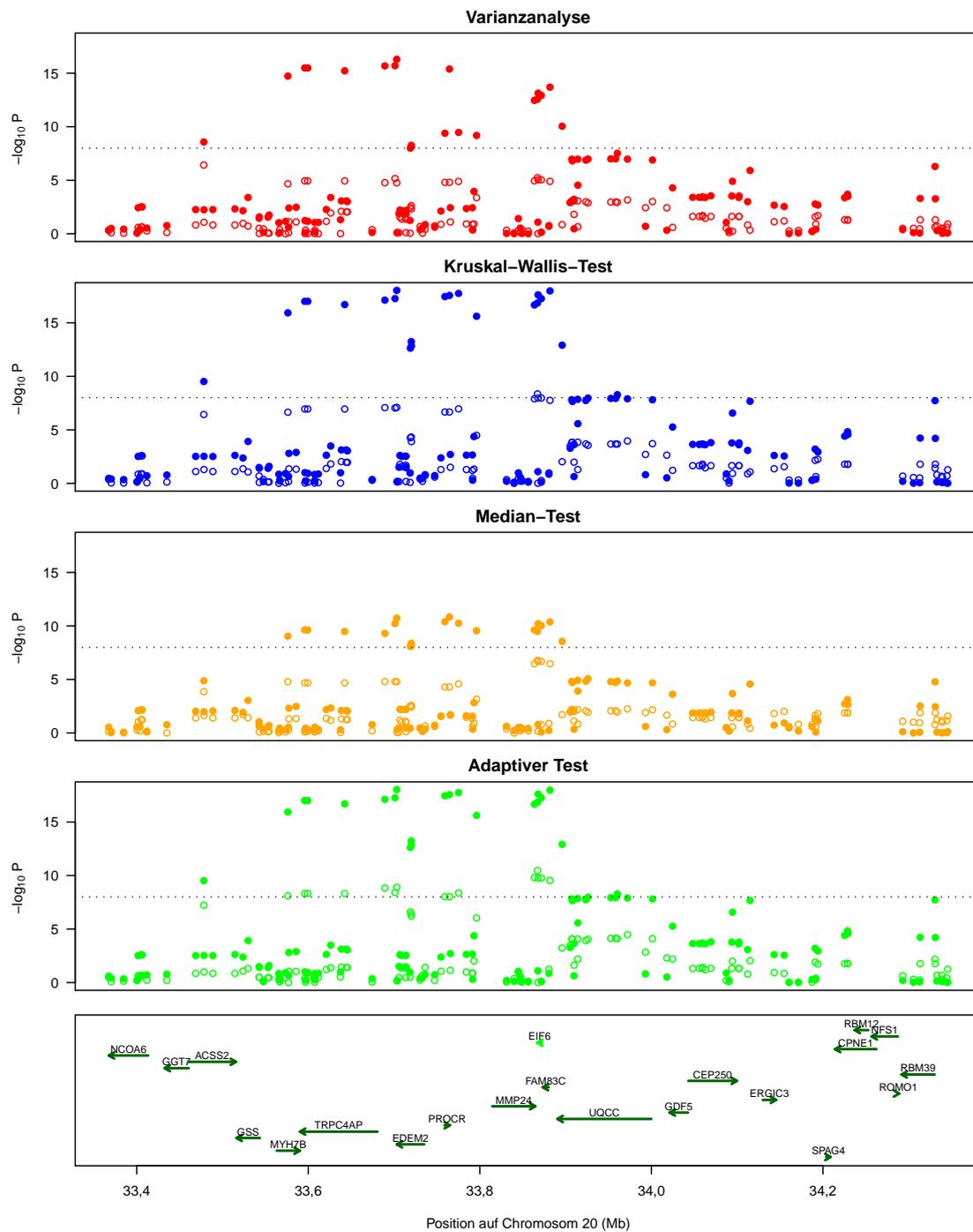


Abbildung 4.2: Regionale Assoziationsabbildung für das Transkript ILMN_1787410 (*EIF6* Gen). Dargestellt sind die negativen logarithmierten p-Werte der Varianzanalyse, des Kruskal-Wallis-Tests, des Median-Tests und des adaptiven Tests für SNPs innerhalb einer Region von 500 kb um das Gen. Die ausgefüllten Kreise stellen die Ergebnisse aus Kohorte 1 dar, während die anderen sich auf Kohorte 2 beziehen. Die gestrichelte Linie gibt das Signifikanzniveau von 10^{-8} an. Im untersten Teil der Abbildung sind die verschiedenen Gene in der Region angegeben, wobei *EIF6* in einem helleren Grün dargestellt ist.

5. Diskussion und Ausblick

Das Ziel dieser Arbeit ist ein statistisches Verfahren für eQTL-Studien mit zwei wichtigen Eigenschaften. Einerseits soll die Methode unter der Nullhypothese das Signifikanzniveau einhalten und andererseits für verschiedene Verteilungssituationen eine möglichst hohe Güte haben. Die durchgeführten Simulationsstudien haben gezeigt, dass klassische Methoden wie die Varianzanalyse und der Kruskal-Wallis-Test nicht geeignet sind, da unterschiedliche Verteilungsformen zu teilweise extremen empirischen Fehlern 1. Art führen. Der Median Test hält zwar in allen untersuchten Simulationsszenarien das Niveau ein, hat aber vor allem bei gleichen Verteilungen eine zu geringe empirische Güte. Als Alternative wird ein optimiertes Klassifikationsschema für einen neuen adaptiven Test vorgeschlagen. Jedoch ist dieses Verfahren bei unterschiedlichen Verteilungsformen zu liberal. Vielversprechend ist eine Kombination der neuen adaptiven Methode mit dem Median Test, die zwar auch nicht immer das Signifikanzniveau einhält, aber zu den kleinsten empirischen Fehlern 1. Art führt.

Idealerweise würde in einer solchen Kombination der Median Test nur dann angewendet, wenn der andere Test zu falsch-positiven Ergebnissen führen würde. Notwendig ist also eine Statistik, die einen Vergleich der Verteilungsformen ermöglicht. Der in dieser Arbeit vorgeschlagene Vergleich der Schätzer für die Schiefe ist jedoch nicht optimal und berücksichtigt außerdem keine anderen Unterschiede in der Verteilungsform. Als Alternativen kommen Statistiken von Omnibus-Tests für allgemeine Alternativhypothesen (s. auch Abschnitt 1.3), wie z.B. der Kolmogorov-Smirnov Test [16] in Frage. Diese Teststatistiken müssten jedoch modifiziert werden, da sie auch auf Lageunterschiede reagieren. Eine andere Möglichkeit ist eine Klassifizierung der Genotyp-spezifischen Verteilungen beispielsweise unter Verwendung des Pearson Systems von Verteilungen [46]. Bei der Entwicklung einer Statistik für den Vergleich der Verteilungsformen sollte

allgemein beachtet werden, dass diese Statistik im Idealfall unabhängig von den verwendeten Teststatistiken ist. Nur in diesem Fall lässt sich analytisch zeigen, dass das adaptive Verfahren robust ist. Für Selektorstatistiken, die auf separaten Stichproben geschätzt werden, kann nur eine asymptotische Verteilungsfreiheit gezeigt werden. Jedoch ist in der hier durchgeführten Simulationsstudie bei großer Fallzahl der Unterschied zu einer Schätzung auf der kombinierten Stichprobe sehr gering.

In dieser Arbeit werden zur Entwicklung eines neuen adaptiven Tests die aus der Literatur bekannten Selektorstatistiken und Gewichte für die lineare Rangstatistik verwendet. Diese verschiedenen Komponenten eines adaptiven Tests können zusätzlich optimiert werden. Bei den Selektorstatistiken ist ein geringer Standardfehler wünschenswert. Dieser könnte beispielsweise durch Verwendung weniger extremer Quantile zur Schätzung der Schiefe und Länge der Verteilungsschwänze erreicht werden. Des Weiteren könnten zusätzliche Gewichte und damit Teststatistiken definiert werden, die möglicherweise in bestimmten Verteilungssituationen Vorteile bieten. Eine Alternative zu den bisher verwendeten festen Grenzen ist die direkte Schätzung von optimalen Gewichtungsfaktoren unter Verwendung der vorliegenden Daten (s. z.B. [6]). Nachteil dieser Methode ist jedoch die längere Rechenzeit für die Dichteschätzung.

Die in dieser Arbeit vorgestellten Simulationsstudien unterscheiden sich vor allem in den verwendeten Verteilungen. In der ersten Studie werden exemplarisch bekannte Verteilungen wie die Normal-, die Laplace- und die χ^2 -Verteilung genutzt, um Unterschiede in den Varianzen, der Schiefe und der Länge der Verteilungsschwänze zu simulieren. Diese Vorgehensweise hat jedoch einen bedeutenden Nachteil, da die genannten Eigenschaften nicht unabhängig voneinander variiert werden können. Beispielsweise hängt bei der χ^2 -Verteilung nicht nur die Schiefe, sondern auch die Varianz und die Länge der Verteilungsschwänze von den Freiheitsgraden ab. Aus diesem Grund wird in der zweiten Simulationsstudie für die Untersuchung der adaptiven Tests die g -und- k -Verteilungsfamilie verwendet, um verschiedene Ausprägungen der Schiefe und der Länge der Verteilungsschwänze untersuchen zu können. In dieser Studie werden keine zusätzlichen Unterschiede in der Varianz berücksichtigt, da keine geschlossene Formel

zur Bestimmung der Varianz bekannt ist. Um diesen Aspekt zu untersuchen, müsste eine andere Verteilungsfamilie verwendet werden.

Bei den in beiden Studien betrachteten Allelhäufigkeiten $p = 0,5$ und $p = 0,3$ waren die Ergebnisse ähnlich. Jedoch wird hiermit ein großer Teil der in GWA-Studien untersuchten genetischen Varianten nicht erfasst, deren Allelhäufigkeit zwischen $p = 0,3$ und $p = 0,05$ liegt. Weitere Simulationsstudien für diese Häufigkeiten sind notwendig, um das asymptotische Verhalten der verschiedenen Tests bei extremeren Unterschieden in der Fallzahl zu untersuchen.

Bei der Anwendung verschiedener Methoden auf Genotyp- und Expressionsdaten der Gutenberg-Herz-Studie werden mit dem neuen kombinierten adaptiven Test mehr bzw. weniger signifikante Assoziationen und Transkripte als mit dem Median-Test bzw. dem Kruskal-Wallis-Test identifiziert. Im Allgemeinen sind die p-Werte aller Methoden jedoch sehr ähnlich. Ein Grund dafür ist sicherlich die hohe Fallzahl in der Gutenberg-Herz-Studie. Selbst bei künstlicher Teilung hat die kleinere Kohorte immer noch mehr als 400 Probanden und ist damit mindestens doppelt so groß wie die Studienpopulationen in den meisten anderen bisher veröffentlichten eQTL-Studien [32]. Ein Vergleich der verschiedenen Verfahren basierend auf solch einer kleineren Studie wäre hilfreich, da sich die Güte der verschiedenen Verfahren bei geringerer Fallzahl stärker unterscheidet. Zusätzlich ist die Variabilität bei der Schätzung der Selektorstatistiken höher. Die p-Werte sollten sich aus diesen Gründen stärker voneinander unterscheiden.

6. Zusammenfassung

Einleitung:

In eQTL-Studien werden SNPs identifiziert, die das Expressionsniveau einzelner Gene beeinflussen. Da die Verteilungen von Expressionsdaten häufig schief sind oder Ausreißer enthalten, sind Standardverfahren wie die Varianzanalyse oder der Kruskal-Wallis-Test nur bedingt geeignet.

Methoden:

Alternativen sind der Median Test und auf Mutual-Information basierende Methoden sowie adaptive Verfahren. Bei letzteren werden Eigenschaften der vorliegenden Verteilungen geschätzt, um einen geeigneten linearen Rangtest auszuwählen. In zwei Simulationsstudien werden die Eigenschaften dieser Tests bei gleichen Verteilungsformen sowie bei Verteilungen mit ungleichen Varianzen, Schiefe in unterschiedlichen Richtungen und variierender Länge der Verteilungsschwänze untersucht.

Ergebnisse:

Nur der Median Test und eine neue Mutual-Information-basierte Methode halten in allen betrachteten Szenarien das Signifikanzniveau ein. Die Varianzanalyse und der Kruskal-Wallis-Test sind bei ungleichen Verteilungsformen teilweise extrem liberal. Das neue Klassifikationsschema bei den adaptiven Tests hat eine höhere empirische Güte als zwei aus der Literatur bekannte Schemata. Die vorgeschlagene Kombination des neuen adaptiven Tests mit dem Median Test führt meistens zu den geringsten empirischen Fehlern, hält jedoch auch nicht immer das Signifikanzniveau ein. In der Gutenberg-Herz-Studie werden mit dem kombinierten adaptiven Test weniger Assoziationen identifiziert als mit dem Kruskal-Wallis-Test, jedoch deutlich mehr als bei Verwendung des Median Tests.

Diskussion:

Eine Kombination der neuen adaptiven Methode mit dem Median-Test ist vielversprechend. Dieser Ansatz könnte mit anderen Entscheidungskriterien für einen der beiden Tests wie Statistiken von Omnibus Tests oder ein Klassifizierungssystem für Verteilungen optimiert werden.

A. Anhang

A.1. Entscheidungsgrenzen für adaptive lineare Rangtests

Tabelle A.1: Adaptiver linearer Rangtest von Büning [15]. Dargestellt sind die Werte für die Selektorstatistiken S_q und T_q , anhand derer ein passender linearer Rangtest ausgewählt wird.

Schiefe	Schwänze	Test
$0,5 \leq \hat{S}_q \leq 2,0$	$1,0 \leq \hat{T}_q < 1,5$	Short-Tails-Test
$0,5 \leq \hat{S}_q \leq 2,0$	$1,5 \leq \hat{T}_q \leq 2,0$	Kruskal-Wallis-Test
$\hat{S}_q > 2,0$	$1,0 \leq \hat{T}_q \leq 2,0$	Right-Skewness-Test
$\hat{S}_q < 0,5$	$1,0 \leq \hat{T}_q \leq 2,0$	Left-Skewness-Test
$\hat{S}_q \geq 0,0$	$\hat{T}_q > 2,0$	Long-Tails-Test

Tabelle A.2: Adaptiver linearer Rangtest von Beier [7]. Dargestellt sind die Werte für die Selektorstatistiken S_q und T_q , anhand derer ein passender linearer Rangtest ausgewählt wird.

Schiefe	Schwänze	Test
$0,5 \leq \hat{S}_q \leq 2,0$	$1,0 \leq \hat{T}_q < 1,5$	Short-Tails-Test
$0,5 \leq \hat{S}_q \leq 2,0$	$1,5 \leq \hat{T}_q \leq 2,0$	Kruskal-Wallis-Test
$0,5 \leq \hat{S}_q \leq 2,0$	$\hat{T}_q > 2,0$	Long-Tails-Test
$\hat{S}_q > 2,0$	$\hat{T}_q \geq 1,0$	Right-Skewness-Test
$\hat{S}_q < 0,5$	$\hat{T}_q \geq 1,0$	Left-Skewness-Test

A.2. Empirische Fehler 1. Art für gleiche Verteilungsformen

A.2.1. Lineare Rangtests

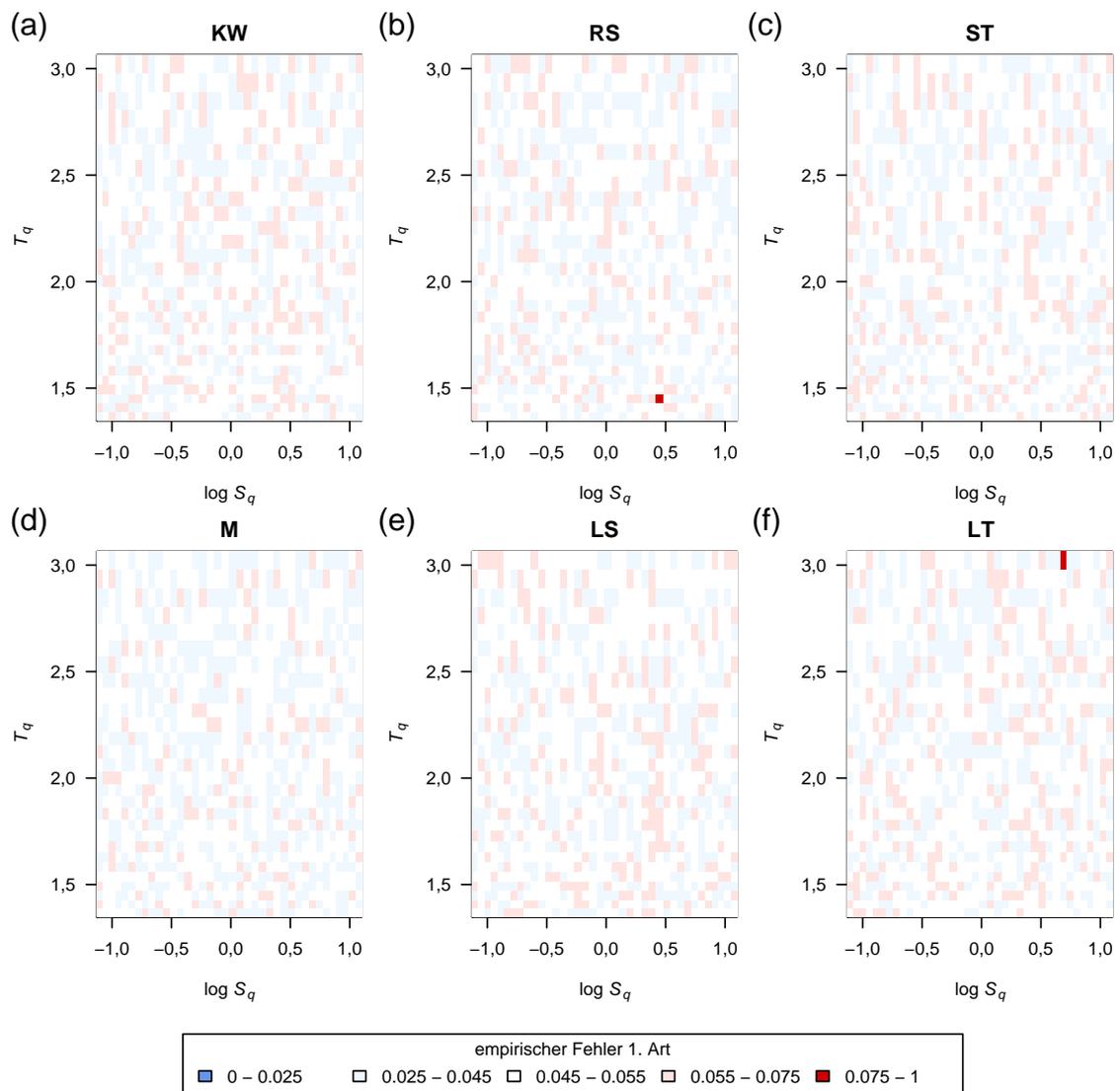


Abbildung A.1: Empirische Fehler 1. Art der linearen Rangtests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,5: (a) Kruskal-Wallis-Test, (b) Right-Skewness-Test, (c) Short-Tails-Test, (d) Median-Test, (e) Left-Skewness-Test, (f) Long-Tails-Test

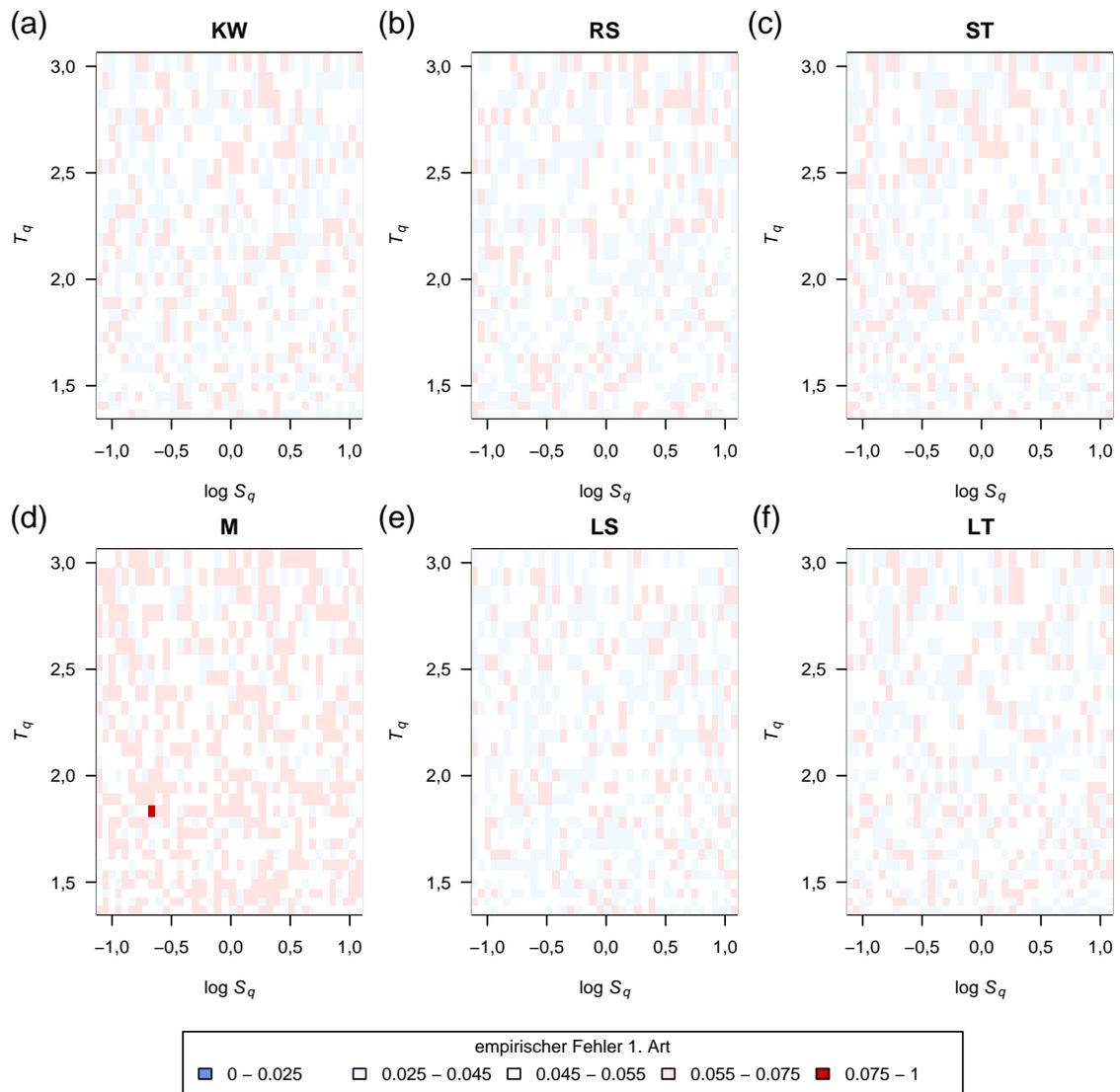


Abbildung A.2: Empirische Fehler 1. Art der linearen Rangtests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,3: (a) Kruskal-Wallis-Test, (b) Right-Skewness-Test, (c) Short-Tails-Test, (d) Median-Test, (e) Left-Skewness-Test, (f) Long-Tails-Test

A.2.2. Adaptive Tests

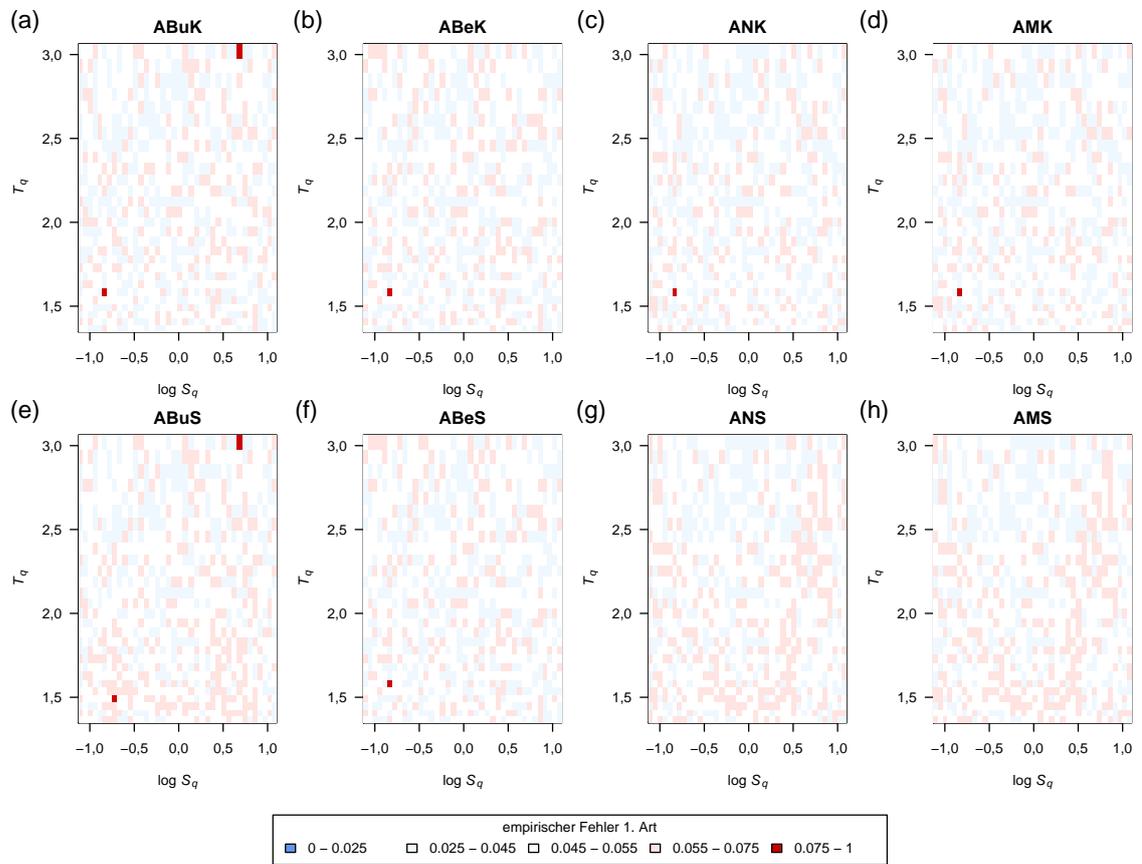


Abbildung A.3: Empirische Fehler 1. Art der adaptiven Tests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,5: (a) Büning kombiniert, (b) Beier kombiniert, (c) Neuer Test kombiniert, (d) Kombiniertes Test kombiniert, (e) Büning separat, (f) Beier separat, (g) Neuer Test separat, (h) Kombiniertes Test separat

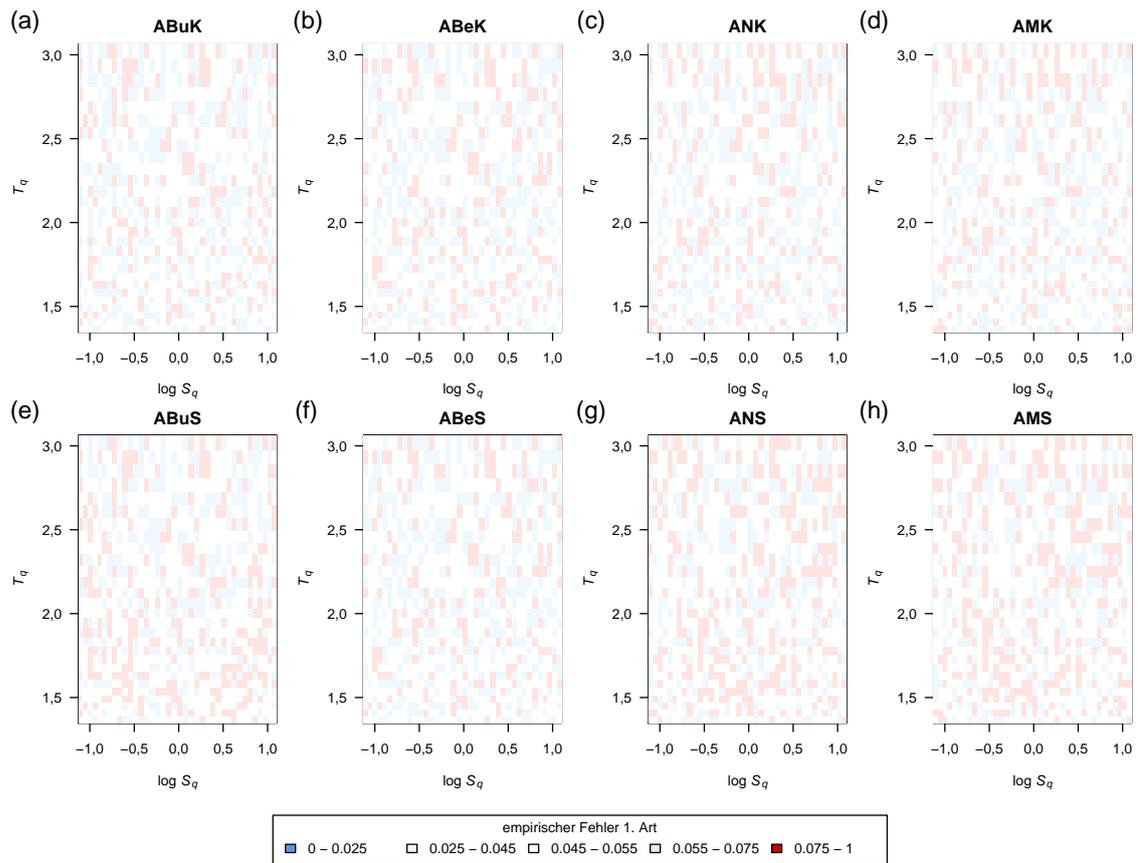


Abbildung A.4: Empirische Fehler 1. Art der adaptiven Tests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz 0,3: (a) Büning kombiniert, (b) Beier kombiniert, (c) Neuer Test kombiniert, (d) Kombiniertes Test kombiniert, (e) Büning separat, (f) Beier separat, (g) Neuer Test separat, (h) Kombiniertes Test separat

A.3. Optimale lineare Rangtests mit neuem Klassifikationsschema

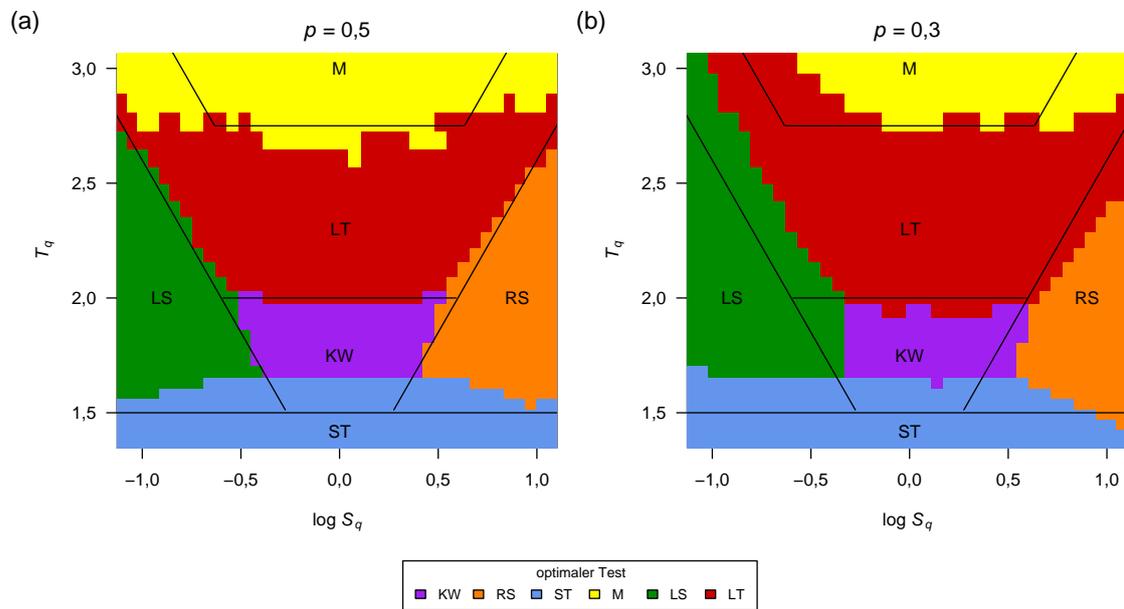


Abbildung A.5: Optimale lineare Rangtests für Verteilungen mit gleicher Verteilungsform und Allelfrequenz (a) 0,5 und (b) 0,3. Dargestellt ist zusätzlich das neue Klassifikationsschema. (KW: Kruskal-Wallis-Test, LS: Left-Skewness-Test, LT: Long-Tails-Test, M: Median-Test, RS: Right-Skewness-Test, ST: Short-Tails-Test)

Literaturverzeichnis

- [1] Alberts R, Terpstra P, Li Y, Breitling R, Nap JP, Jansen RC: Sequence polymorphisms cause many false cis eQTLs. *PLoS ONE* 2, e622 (2007)
- [2] Archer KJ, Reese SE: Detection call algorithms for high-throughput gene expression microarray data. *Brief Bioinform* 11, 244–252 (2009)
- [3] Balanda KP, MacGillivray HL: Kurtosis: A critical review. *Am Stat* 42, 111–119 (1988)
- [4] Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, Darot JFJ, Ritchie ME, Lynch AG, Tavaré S: A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res* 38, e17 (2009)
- [5] Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: Reverse engineering of regulatory networks in human B cells. *Nat Genet* 37, 382–390 (2005)
- [6] Behnen K, Neuhaus G: Rank Tests with Estimated Scores and their Application. B.G. Teubner, Stuttgart (1989)
- [7] Beier F: Adaptive Tests bei nicht-monotonen Dosis-Wirkungsbeziehungen. Diss., Technische Universität Dortmund (1996)
- [8] Bickel PS: On some robust estimates of location. *Ann Math Stat* 36, 847–858 (1965)

-
- [9] Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193 (2003)
- [10] Bradley JV: Robustness? *Br J Math Stat Psychol* 31, 144–152 (1978)
- [11] Brown GW, Mood AM: On median tests for linear hypotheses. In: Neyman J (Hrsg) *Second Berkeley Symposium on Mathematical Statistics and Probability*, 159–166: University of California Press, Berkeley (1951)
- [12] Büning H: *Robuste und Adaptive Tests*. de Gruyter, Berlin (1991)
- [13] Büning H: Adaptive tests for the c-sample location problem - the case of two sided alternatives. *Commun Stat Theory* 25, 1569–1582 (1996)
- [14] Büning H: Robust analysis of variance. *J Appl Stat* 24, 319–332 (1997)
- [15] Büning H: Adaptive tests for the c-sample location problem. In: Schipp B, Krämer W (Hrsg) *Statistical Inference, Econometric Analysis and Matrix Algebra*, 3–17: Physica-Verlag, Heidelberg (2009)
- [16] Büning H, Trenkler G: *Nichtparametrische Statistische Methoden*. de Gruyter, New York (1994)
- [17] Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS: Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet* 33, 422–425 (2003)
- [18] Cover TM, Thomas JA: *Elements of Information Theory*. Wiley, New York (1991)
- [19] Dawy Z, Göbel B, Hagenauer J, Andreoli C, Meitinger T, Müller JC: Gene mapping and marker clustering using Shannon's mutual information. *IEEE/ACM Trans Comput Biol Bioinform* 3, 47–56 (2006)

-
- [20] Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Arcelus MG, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis ET, Antonarakis SE: Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325, 1246–1250 (2009)
- [21] Doksum AK: Location and asymmetry. *Scand J Stat* 2, 11–22 (1975)
- [22] Du P, Kibbe WA, Lin SM: lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 24, 1547–1548 (2008)
- [23] Duan S, Huang RS, Zhang W, Bleibel WK, Roe CA, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ, Dolan ME: Genetic architecture of transcript-level variation in humans. *Am J Hum Genet* 82, 1101–1113 (2008)
- [24] Dunning MJ, Smith ML, Ritchie ME, Tavaré S: beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics* 23, 2183–2184 (2007)
- [25] Embrechts P, Mikosch T, Klüppelberg C: Modelling Extremal Events. Springer, Berlin (1997)
- [26] Fagerland MW, Sandvik L: The Wilcoxon-Mann-Whitney test under scrutiny. *Stat Med* 28, 1487–1497 (2009)
- [27] Freidlin B, Gastwirth JL: Should the median test be retired from general use? *Am Statist* 54, 161–164 (2000)
- [28] Gastwirth JL: Percentile modifications of two sample rank tests. *J Am Stat Assoc* 60, 1127–1141 (1965)
- [29] Glass GV, Peckham PD, Sanders JR: Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Rev Educ Res* 42, 237–288 (1972)
- [30] Göbel B, Dawy Z, Hagenauer J, Müller JC: An approximation to the distri-

- bution of finite sample size mutual information estimates. In: Proceedings IEEE International Conference on Communications, Bd. 2, 1102–1106 (2005)
- [31] Göring HHH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JBM, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J: Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* 39, 1208–1216 (2007)
- [32] Grundberg E, Kwan T, Pastinen TM: Analysis of the impact of genetic variation on human gene expression. *Methods Mol Biol* 628, 321–339 (2010)
- [33] Handl A: Masszahlen zur Klassifizierung von Verteilungen bei der Konstruktion adaptiver verteilungsfreier Tests im unverbundenen Zweistichproben-Problem. Diss., Freie Universität Berlin (1985)
- [34] Haynes MA, MacGillivray HL, Mengersen KL: Robustness of ranking and selection rules using generalised g -and- k distributions. *J Stat Plan Inference* 65, 45–66 (1997)
- [35] Heap G, Trynka G, Jansen R, Bruinenberg M, Schwertz M, Dinesen L, Hunt K, Wijmenga C, van Heel D, Franke L: Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med Genomics* 2, 1 (2009)
- [36] Hill NJ, Padmanabhan AR, Puri ML: Adaptive nonparametric procedures and applications. *J R Stat Soc Ser C Appl Stat* 37, 205–218 (1988)
- [37] Hindorff LA, Junkins HA, Hall PN, Mehta JP, Manolio TA: A catalog of published genome-wide association studies. (Tag des Zugriffs: 23.09.2010) www.genome.gov/gwastudies/
- [38] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-

- wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 106, 9362–9367 (2009)
- [39] Hoaglin DC: Summarizing shape numerically: The g-and-h distributions. In: Hoaglin DC, Mosteller F, Tukey JW (Hrsg) *Exploring Data Tables, Trends, and Shapes*, Kap. 11, 461–514: Wiley, New York (1985)
- [40] Hogg RV: Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *J Am Stat Assoc* 69, 909–923 (1974)
- [41] Hogg RV, Fisher DM, Randles RH: A two sample adaptive distribution-free test. *J Am Stat Assoc* 70, 656–661 (1975)
- [42] Hogg RV, Randles RH: Adaptive distribution-free regression methods and their applications. *Technometrics* 17, 399–407 (1975)
- [43] Hong H, Su Z, Ge W, Shi L, Perkins R, Fang H, Xu J, Chen JJ, Han T, Kaput J, Fuscoe JC, Tong W: Assessing batch effects of genotype calling algorithm BRLMM for the Affymetrix GeneChip Human Mapping 500 K array set using 270 HapMap samples. *BMC Bioinformatics* 9, S17 (2008)
- [44] Hunt KA, Zhernakova A, Turner G, Heap GAR, Franke L, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, Panesar D, Gwilliam R, Takeuchi F, McLaren WM, Holmes GKT, Howdle PD, Walters JRF, Sanders DS, Playford RJ, Trynka G, Mulder CJJ, Mearin ML, Verbeek WHM, Trimble V, Stevens FM, O’Morain C, Kennedy NP, Kelleher D, Pennington DJ, Strachan DP, McArdle WL, Mein CA, Wapenaar MC, Deloukas P, McGinnis R, McManus R, Wijmenga C, van Heel DA: Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet* 40, 395–402 (2008)
- [45] Jarinova O, Stewart AFR, Roberts R, Wells G, Lau P, Naing T, Buerki C, McLean BW, Cook RC, Parker JS, McPherson R: Functional analysis of the chromosome 9p21.3 coronary artery disease risk locus. *Arterioscler Thromb Vasc Biol* 29, 1671–1677 (2009)

-
- [46] Johnson NL, Kotz S, Balakrishnan N: Continuous Univariate Distributions, Bd. 2. Wiley, New York, 2. Aufl. (1995)
- [47] Katz S, Irizarry RA, Lin X, Tripputi M, Porter MW: A summarization approach for Affymetrix GeneChip data using a reference training set from a large, biologically diverse database. *BMC Bioinformatics* 7, 464 (2006)
- [48] Khan A, Rayner GD: Robustness to non-normality of common tests for the many-sample location problem. *J Appl Math Decis Sci* 7, 187–206 (2003)
- [49] Kim TH, White H: On more robust estimation of skewness and kurtosis: Simulation and application to the S&P500 index. Economics Working Paper Series 2003-12, Department of Economics, University of California San Diego (2004)
- [50] Kochi Y, Okada Y, Suzuki A, Ikari K, Terao C, Takahashi A, Yamazaki K, Hosono N, Myouzen K, Tsunoda T, Kamatani N, Furuichi T, Ikegawa S, Ohmura K, Mimori T, Matsuda F, Iwamoto T, Momohara S, Yamanaka H, Yamada R, Kubo M, Nakamura Y, Yamamoto K: A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. *Nat Genet* 42, 515–519 (2010)
- [51] Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D: Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 40, 1253–1260 (2008)
- [52] Kössler W: Some c-sample rank tests of homogeneity against ordered alternatives based on U-statistics. *J Nonparametr Stat* 17, 777–795 (2005)
- [53] Kössler W: Nonparametric Location Tests Against Restricted Alternatives. Shaker Verlag, Aachen (2006)
- [54] Kössler W, Kumar N: An adaptive test for the two-sample location problem

- based on U-statistics. *Commun Stat Simulat* 37, 1329–1346 (2008)
- [55] Kristensen VN, Edvardsen H, Tsalenko A, Nordgard SH, Sørлие T, Sharan R, Vailaya A, Ben-Dor A, Lønning PE, Lien S, Omholt S, Syvänen AC, Yakhini Z, Børresen-Dale AL: Genetic variation in putative regulatory loci controlling gene expression in breast cancer. *Proc Natl Acad Sci U S A* 103, 7735–7740 (2006)
- [56] Kristensen VN, Tsalenko A, Geisler J, Faldaas A, Grenaker GI, Lingjaerde OC, Fjeldstad S, Yakhini Z, Lønning PE, Børresen-Dale AL: Multilocus analysis of SNP and metabolic data within a given pathway. *BMC Genomics* 7, 5 (2006)
- [57] Kruskal WH: A nonparametric test for the several sample problem. *Ann Math Stat* 23, 525–540 (1952)
- [58] Kuhn K, Baker SC, Chudin E, Lieu MH, Oeser S, Bennett H, Rigault P, Barker D, McDaniel TK, Chee MS: A novel, high-performance random array platform for quantitative gene expression profiling. *Genome Res* 14, 2347–2356 (2004)
- [59] Lehmann E: Elements of Large-Sample Theory. Springer, New York (1999)
- [60] Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, Vermeire S, Dewit O, de Vos M, Dixon A, Demarche B, Gut I, Heath S, Foglio M, Liang L, Laukens D, Mni M, Zelenika D, Van Gossum A, Rutgeerts P, Belaiche J, Lathrop M, Georges M: Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet* 3, e58 (2007)
- [61] Lin SM, Du P, Huber W, Kibbe WA: Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res* 36, e11 (2008)
- [62] Liu X, Krishnan A, Mondry A: An entropy-based gene selection method for cancer classification using microarray data. *BMC Bioinformatics* 6, 76 (2005)

-
- [63] Mahoney M, Magel R: Estimation of the power of the Kruskal-Wallis test. *Biom J* 38, 613–630 (1996)
- [64] Manolio TA, Collins FS: The HapMap and genome-wide association studies in diagnosis and therapy. *Annu Rev Med* 60, 443–456 (2009)
- [65] Martinez J, Iglewicz B: Some properties of the tukey g and h family of distributions. *Commun Stat Theory* 13, 353–369 (1984)
- [66] The International HapMap Consortium: A second generation human haplotype map of over 31 million SNPs. *Nature* 449, 851–861 (2007)
- [67] Tobacco and Genetics Consortium: Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 42, 441–447 (2010)
- [68] McDougall MK, Rayner GD: Robustness to non-normality of various tests for the one-sample location problem. *J Appl Math Decis Sci* 8, 235–246 (2004)
- [69] Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E, Heinzmann A, Simma B, Frischer T, Willis-Owen SAG, Wong KCC, Illig T, Vogelberg C, Weiland SK, von Mutius E, Abecasis GR, Farrall M, Gut IG, Lathrop GM, Cookson WOC: Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 448, 470–473 (2007)
- [70] Mood AM, Graybill FA, Boes DC: Introduction to the Theory of Statistics. McGraw-Hill, Singapore (1974)
- [71] NCI-NHGRI Working Group on Replication in Association Studies, Chanoock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni JF, Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi

- C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS: Replicating genotype-phenotype associations. *Nature* 447, 655–660 (2007)
- [72] Neuhäuser M, Büning H, Hothorn LA: Maximum test versus adaptive tests for the two-sample location problem. *J Appl Stat* 31, 215–227 (2004)
- [73] Pasmant E, Laurendeau I, Héron D, Vidaud M, Vidaud D, Bièche I: Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an anti-sense noncoding RNA whose expression coclusters with ARF. *Cancer Res* 67, 3963–3969 (2007)
- [74] Pratt JW: Robustness of some procedures for the two-sample location problem. *J Am Stat Assoc* 59, 665–680 (1964)
- [75] Priness I, Maimon O, Ben-Gal I: Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics* 8, 111 (2007)
- [76] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81, 559–575 (2007)
- [77] Puri ML: Some aspects of nonparametric inference. *Int Stat Rev* 40, 299–327 (1972)
- [78] Rayner GD, MacGillivray HL: Weighted quantile-based estimation for a class of transformation distributions. *Comput Stat Data Anal* 39, 401–433 (2002)
- [79] Roussas GG: A First Course in Mathematical Statistics. Addison-Wesley, Reading (1973)
- [80] Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, Barrett JH, König IR, Stevens

- SE, Szymczak S, Tregouet DA, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Braenne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR, Schunkert H, C WTCC, the Cardiogenics Consortium: Genomewide association analysis of coronary artery disease. *N Engl J Med* 357, 443–453 (2007)
- [81] Sandhu MS, Waterworth DM, Debenham SL, Wheeler E, Papadakis K, Zhao JH, Song K, Yuan X, Johnson T, Ashford S, Inouye M, Luben R, Sims M, Hadley D, McArdle W, Barter P, Kesäniemi YA, Mahley RW, McPherson R, Grundy SM, Consortium WTCC, Bingham SA, Khaw KT, Loos RJF, Waeber G, Barroso I, Strachan DP, Deloukas P, Vollenweider P, Wareham NJ, Mooser V: LDL-cholesterol concentrations: a genome-wide association study. *Lancet* 371, 483–491 (2008)
- [82] Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusk AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich R: Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* 6, e107 (2008)
- [83] Shannon CE: A mathematical theory of communication. *Bell Syst Tech J* 27, 379–423 (1948)
- [84] Shoemaker LH, Rosenberger JL: Moments and efficiency of the median and trimmed mean for finite populations. *Commun Stat Simulat* 12, 411–422 (1983)
- [85] Smith NL, Chen MH, Dehghan A, Strachan DP, Basu S, Soranzo N, Hayward C, Rudan I, Sabater-Lleal M, Bis JC, de Maat MPM, Rumley A, Kong X, Yang Q, Williams FMK, Vitart V, Campbell H, Mälarstig A, Wiggins KL, Duijn CMV, McArdle WL, Pankow JS, Johnson AD, Silveira A, McKnight B, Uitterlinden AG, Consortium; WTCC, Aleksic N, Meigs JB, Peters A, Koenig W, Cushman M, Kathiresan S, Rotter JL, Bovill EG, Hofman A, Boer-

- winkle E, Tofler GH, Peden JF, Psaty BM, Leebeek F, Folsom AR, Larson MG, Spector TD, Wright AF, Wilson JF, Hamsten A, Lumley T, Wittteman JCM, Tang W, O'Donnell CJ: Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. *Circulation* 121, 1382–1392 (2010)
- [86] Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FAS, Zhernakova A, Hinks A, Guiducci C, Chen R, Alfredsson L, Amos CI, Ardlie KG, Consortium BIRAC, Barton A, Bowes J, Brouwer E, Burtt NP, Catanese JJ, Coblyn J, Coenen MJH, Costenbader KH, Criswell LA, Crusius JBA, Cui J, de Bakker PIW, Jager PLD, Ding B, Emery P, Flynn E, Harrison P, Hocking LJ, Huizinga TWJ, Kastner DL, Ke X, Lee AT, Liu X, Martin P, Morgan AW, Padyukov L, Posthumus MD, Radstake TRDJ, Reid DM, Seielstad M, Seldin MF, Shadick NA, Steer S, Tak PP, Thomson W, van der Helm-van Mil AHM, van der Horst-Bruinsma IE, van der Schoot CE, van Riel PLCM, Weinblatt ME, Wilson AG, Wolbink GJ, Wordsworth BP, Consortium YEAR, Wijmenga C, Karlson EW, Toes REM, de Vries N, Begovich AB, Worthington J, Siminovitch KA, Gregersen PK, Klareskog L, Plenge RM: Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* 42, 508–514 (2010)
- [87] Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavaré S, Deloukas P, Dermitzakis ET: Population genomics of human gene expression. *Nat Genet* 39, 1217–1224 (2007)
- [88] Suchindran S, Rivedal D, Guyton JR, Milledge T, Gao X, Benjamin A, Rowell J, Ginsburg GS, McCarthy JJ: Genome-wide association study of Lp-PLA(2) activity and mass in the Framingham Heart Study. *PLoS Genet* 6, e1000928 (2010)
- [89] Szymczak S, Igl BW, Ziegler A: Detecting SNP-expression associations: A comparison of mutual information and median test with standard statistical approaches. *Stat Med* 28, 3581–3596 (2009)

- [90] Szymczak S, Nuzzo A, Fuchsberger C, Schwarz DF, Ziegler A, Bellazi R, Igl BW: Genetic association studies for gene expressions: Permutation-based mutual information in a comparison with standard ANOVA and as a novel approach for feature selection. *BMC Proc* 1, S9 (2007)
- [91] Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ, Johansen CT, Fouchier SW, Isaacs A, Peloso GM, Barbalic M, Ricketts SL, Bis JC, Aulchenko YS, Thorleifsson G, Feitosa MF, Chambers J, Orho-Melander M, Melander O, Johnson T, Li X, Guo X, Li M, Cho YS, Go MJ, Kim YJ, Lee JY, Park T, Kim K, Sim X, Ong RTH, Croteau-Chonka DC, Lange LA, Smith JD, Song K, Zhao JH, Yuan X, Luan J, Lamina C, Ziegler A, Zhang W, Zee RYL, Wright AF, Witteman JCM, Wilson JF, Willemssen G, Wichmann HE, Whitfield JB, Waterworth DM, Wareham NJ, Waeber G, Vollenweider P, Voight BF, Vitart V, Uitterlinden AG, Uda M, Tuomilehto J, Thompson JR, Tanaka T, Surakka I, Stringham HM, Spector TD, Soranzo N, Smit JH, Sinisalo J, Silander K, Sijbrands EJG, Scuteri A, Scott J, Schlessinger D, Sanna S, Salomaa V, Saharinen J, Sabatti C, Ruukonen A, Rudan I, Rose LM, Roberts R, Rieder M, Psaty BM, Pramstaller PP, Pichler I, Perola M, Penninx BWJH, Pedersen NL: Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466, 707–713 (2010)
- [92] Tsalenko A, Sharan R, Kristensen V, Edvardsen H, Børresen-Dale AL, Bendor A, Yakhini Z: Analysis of SNP-expression association matrices. *J Bioinform Comput Biol* 4, 259–274 (2006)
- [93] Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segrè AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonycastle LL, Boström KB, Bravenboer B, Bumpstead S, Burt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney ASF, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder

- C, Isomaa B, Jackson AU, Johnson PRV, Jørgensen T, Kao WHL, Klopp N, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieverse A, Lindgren CM, Lyssenko V, Marre M, Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Perry JRB, Petersen AK, Platou C, Proença C, Prokopenko I, Rathmann W, Rayner NW, Robertson NR, Rochelea: Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 42, 579–589 (2010)
- [94] Wallace C, Newhouse SJ, Braund P, Zhang F, Tobin M, Falchi M, Ahmadi K, Dobson RJ, Marçano ACB, Hajat C, Burton P, Deloukas P, Brown M, Connell JM, Dominiczak A, Lathrop GM, Webster J, Farrall M, Spector T, Samani NJ, Caulfield MJ, Munroe PB: Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am J Hum Genet* 82, 139–149 (2008)
- [95] Weale ME: Quality control for genome-wide association studies. *Methods Mol Biol* 628, 341–372 (2010)
- [96] Weber C, Zerneck A, Libby P: The multifaceted contributions of leukocyte subsets to atherosclerosis: lessons from mouse models. *Nat Rev Immunol* 8, 802–815 (2008)
- [97] Webster JA, Gibbs JR, Clarke J, Ray M, Zhang W, Holmans P, Rohrer K, Zhao A, Marlowe L, Kaleem M, McCorquodale DS, Cuello C, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, Coon KD, Craig DW, Pearson JV, NACC-Neuropathology Group, Heward CB, Reiman EM, Stephan D, Hardy J, Myers AJ: Genetic control of human brain transcript expression in Alzheimer disease. *Am J Hum Genet* 84, 445–458 (2009)
- [98] Welch BL: On the comparison of several mean values: an alternative approach. *Biometrika* 38, 330–336 (1951)
- [99] Wells CS, Hintze JM: Dealing with assumptions underlying statistical tests. *Psychol Sch* 44, 495–502 (2007)

- [100] Wilcoxon RR, Charlin VL: Comparing medians: A Monte Carlo study. *J Educ Behav Stat* 11, 263–274 (1986)
- [101] Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, Strait J, Duren WL, Maschio A, Busonero F, Mulas A, Albai G, Swift AJ, Morken MA, Narisu N, Bennett D, Parish S, Shen H, Galan P, Meneton P, Hercberg S, Zelenika D, Chen WM, Li Y, Scott LJ, Scheet PA, Sundvall J, Watanabe RM, Nagaraja R, Ebrahim S, Lawlor DA, Ben-Shlomo Y, Davey-Smith G, Shuldiner AR, Collins R, Bergman RN, Uda M, Tuomilehto J, Cao A, Collins FS, Lakatta E, Lathrop GM, Boehnke M, Schlessinger D, Mohlke KL, Abecasis GR: Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* 40, 161–169 (2008)
- [102] Wu X, Jin L, Xiong M: Mutual information for testing gene-environment interaction. *PLoS ONE* 4, e4578 (2009)
- [103] Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, Rossmann H, Eleftheriadis M, Sinning CR, Schnabel RB, Lubos E, Mennerich D, Rust W, Perret C, Proust C, Nicaud V, Loscalzo J, Hübner N, Tregouet D, Münzel T, Ziegler A, Tiret L, Blankenberg S, Cambien F: Genetics and beyond - the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* 5, e1069 (2010)
- [104] Zenga M: Kurtosis. In: Kotz S, Read CB, Balakrishanan N, Vidakovic B (Hrsg) *Encyclopedia of Statistical Sciences*: Wiley, New York, 2. Aufl. (2006)
- [105] Ziegler A, König IR, Thompson JR: Biostatistical aspects of genome-wide association studies. *Biom J* 50, 8–28 (2008)
- [106] Zimmerman DW: Statistical significance levels of nonparametric tests biased by heterogeneous variances of treatment groups. *J Gen Psychol* 127, 354–364 (2000)

Danksagung

An erster Stelle richtet sich mein Dank an Herrn Univ.-Prof. Dr. Andreas Ziegler für die umfassende wissenschaftliche Betreuung und die Möglichkeit, neben der Dissertation an verschiedenen spannenden Forschungsprojekten mitzuarbeiten. Besonders bedanken will ich mich auch für die Freiheit, das Dissertationsthema nach meinen Interessen auszuwählen.

Herrn Dr. Bernd-Wolfgang Igl möchte ich für die vielen hilfreichen und kreativen Diskussionen sowie zahlreichen fachlichen Hinweise danken.

Bei Herrn Univ.-Prof. Dr. Stefan Blankenberg bedanke ich mich für die gute Zusammenarbeit und die Möglichkeit, die Genotyp- und Expressionsdaten aus der Gutenberg-Herz-Studie für meine Dissertation zu verwenden. Bedanken möchte ich mich darüber hinaus bei Frau Dr. Tanja Zeller für die hilfreichen Hintergrund-Informationen zu den labortechnischen Aspekten bei der Durchführung der Microarray-Experimente.

Ein großes Dankeschön geht an alle Mitarbeiterinnen und Mitarbeiter des Instituts für Medizinische Biometrie und Statistik für die hervorragende Zusammenarbeit und Hilfestellung bei allen möglichen Problemen. Besonders möchte ich mich bei Herrn M. Sc. Markus Scheinhardt für die technische Unterstützung bei der Implementierung der adaptiven Tests und der Simulationsstudien bedanken. Außerdem danke ich ihm und Frau M. Sc. Maren Vens für die Fehlerkorrektur und die zahlreichen Diskussionen.

Mein Dank gebührt außerdem meinen Eltern Elfriede und Manfred Szymczak sowie meiner Schwester Sonja Szymczak für die vielen Anregungen zu stilistischen und orthografischen Überarbeitungen.