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

Statistische Verfahren zur Analyse von Assoziationen seltener genetischer Varianten mit komplexen Erkrankungen: Ein Vergleich verschiedener Gruppierungsmethoden

Inauguraldissertation
zur
Erlangung der Doktorwürde
der Universität zu Lübeck

– Aus der Sektion Medizin –

vorgelegt von

Carmen Dering

aus Stralsund

Lübeck, 2. September 2016

1. Berichterstatter:	Prof. Dr. rer. nat. Andreas Ziegler
2. Berichterstatter:	Dr. phil. Dipl.-Psych. Hans-Jürgen Rumpf

Tag der mündlichen Prüfung: 23.08.2016

Zum Druck genehmigt.
Lübeck,
Promotionskommission der Sektion Medizin

Inhaltsverzeichnis

Abbildungsverzeichnis v

Tabellenverzeichnis vii

1 Einleitung 1

1.1 Hintergrund . . . 1
1.2 Bisherige Vergleichsarbeiten zu Gruppierungsmethoden . . . 5
1.3 Ziele der Arbeit . . . 7
1.4 Aufbau der Arbeit . . . 10

2 Material und Methoden 12

2.1 Datenvorverarbeitung: Bildung einer Region von Interesse . . . 12
2.1.1 Annotationsdatenbanken . . . 14
2.1.2 Filterkriterien . . . 16
2.1.2.1 Minor Allele Frequency (MAF) . . . 16
2.1.2.2 Funktionalität von Varianten . . . 18
2.1.3 Die einfache Region von Interesse . . . 19
2.1.4 Die gefilterte Region von Interesse . . . 21
2.2 Eigenschaften von Gruppierungsmethoden . . . 22
2.2.1 Kodierung der Region von Interesse . . . 22
2.2.1.1 Indikatorkodierung . . . 23
2.2.1.2 Genotypkodierung . . . 23
2.2.2 Putative Funktion von Varianten . . . 24
2.2.2.1 Fehlklassifizierung . . . 25
2.2.2.2 Lösungsansätze bei Fehlklassifizierungen . . . 25
2.2.3 Gemeinsame Berücksichtigung seltener und häufiger Varianten 26
2.2.4 Regressionsansatz . . . 28
2.2.4.1 Allgemeines Regressionsmodell . . . 29

2.2.4.2 Art des Phänotyps 30
2.2.4.3 Kovariablen 31
2.2.4.4 Gewichte 32
2.2.5 Struktur der Teststatistik 34
2.2.5.1 Lineare Teststatistiken 34
2.2.5.2 Quadratische Teststatistiken 35
2.2.5.3 Vergleich linearer und quadratischer Teststatistiken . 36
2.2.6 Schätzung des p-Werts 37
2.3 Die untersuchten Gruppierungsmethoden 39
2.3.1 CAST – Cohort Allelic Sum Test 42
2.3.2 CMC – Combined and Multivariate Collapsing 43
2.3.3 RC – Rarecover 44
2.3.4 RVT1 – Rare Variant Test 1 46
2.3.5 RVT2 – Rare Variant Test 2 48
2.3.6 WSS – Weighted Sum Statistic 49
2.3.7 CMAT – Cumulative Minor–Allele Test 52
2.3.8 VT – Variable Threshold 53
2.3.9 ASUM – Adaptive Summation Test 56
2.3.10 KBAC – Kernel Based Adaptive Cluster 58
2.3.11 C-α – C-alpha-based Test 61
2.3.12 FPCA – Functional Principal Component Analysis 62
2.3.13 PWST – P-value Weighted Sum Test 66
2.3.14 SKAT – Sequencing Kernel Association Test 68
2.3.15 SKAT-O - Optimal Unified Sequencing Kernel Association Test 71

3 Simulationsstudie 74
3.1 Simulationsaufbau 75
3.2 Filterszenarien – Die Regionen von Interesse 77
3.3 Phänotypen und Kovariablen 78
3.4 Schätzung des p-Werts durch Permutation 78
3.5 Schätzung des Fehlers 1. Art und der Teststärke 80
3.5.1 Der empirische Fehler 1. Art 80
3.5.2 Die empirische Teststärke 81
3.5.3 Die minimale Teststärke 81
3.5.4 Das Signifikanzniveau α 82
3.6 Ergebnisse 82

3.6.1 Binärer Phänotyp ohne Kovariablen 83
3.6.2 Binärer Phänotyp mit Kovariablen 88
3.6.3 Quantitativer Phänotyp ohne Kovariablen 92
3.6.4 Quantitativer Phänotyp mit Kovariablen 100
3.7 Interpretation 102

4 Anwendung: Das Gen *SLCO1B1* bei Leukämie 106
4.1 Einführung 106
4.2 Filterszenarien – Die Regionen von Interesse 107
4.3 Ergebnisse 109
4.4 Interpretation 110

5 Diskussion und Ausblick 112

6 Zusammenfassung 121

Literatur 123

Danksagung

Lebenslauf

Publikationsverzeichnis

Abbildungsverzeichnis

2.1 Flussdiagramm zur Bildung einer Region von Interesse 13
2.2 Bildung der einfachen Region von Interesse 20
2.3 Bildung der gefilterten Region von Interesse 21

3.1 Quantil-Quantil-Plots der p-Wert-Verteilung. Fall-Kontroll-Status ohne Kovariablen. Gen-weise Gruppierung der Varianten mit einer Frequenz des seltenen Allels $\leq$ 0,01. 85
3.2 Quantil-Quantil-Plots der p-Wert-Verteilung. Fall-Kontroll-Status ohne Kovariablen. Gen-weise Gruppierung der Varianten mit einer Frequenz des seltenen Allels $\leq$ 0,05. 86
3.3 Streudiagramme der assoziierten Regionen von Interesse bzgl. des Anteils der verursachenden Varianten und der empirischen Teststärke. Betrachtung des Fall-Kontroll-Status ohne Kovariablen. Gen-weise Gruppierung der Varianten mit einer Frequenz des seltenen Allels $\leq$ 0,01. 89
3.4 Streudiagramme der assoziierten Regionen von Interesse bzgl. des Anteils der verursachenden Varianten und der empirischen Teststärke. Betrachtung des Fall-Kontroll-Status ohne Kovariablen. Gen-weise Gruppierung der Varianten mit einer Frequenz des seltenen Allels $\leq$ 0,05. 90
3.5 Quantil-Quantil-Plots der p-Wert-Verteilung. Betrachtung des Fall-Kontroll-Status mit Kovariablen. Gen-weise Gruppierung der Varianten mit einer Frequenz des seltenen Allels $\leq$ 0,01 und $\leq$ 0,05 93

3.6 Streudiagramme der assoziierten Region von Interesse bzgl. des Anteils der verursachenden Varianten und der empirischen Teststärke. Betrachtung des Fall-Kontroll-Status mit Kovariablen. Gen-weise Gruppierung der Varianten mit einer Frequenz des seltenen Allels $\leq 0{,}01$ und $\leq 0{,}05$ 94
3.7 Quantil-Quantil-Plots der p-Wert-Verteilung. Betrachtung des quantitativen Phänotyps ohne Kovariablen. Gen-weise Gruppierung der Varianten mit einer Frequenz des seltenen Allels $\leq 0{,}01$. 95
3.8 Quantil-Quantil-Plots der p-Wert-Verteilung. Betrachtung des quantitativen Phänotyps ohne Kovariablen. Gen-weise Gruppierung der Varianten mit einer Frequenz des seltenen Allels $\leq 0{,}05$. 96
3.9 Streudiagramme der assoziierten Regionen von Interesse bzgl. des Anteils der verursachenden Varianten und der empirischen Teststärke. Betrachtung des quantitativen Phänotyps ohne Kovariablen. Gen-weise Gruppierung der Varianten mit einer Frequenz des seltenen Allels $\leq 0{,}01$. 97
3.10 Streudiagramme der assoziierten Regionen von Interesse bzgl. des Anteils der verursachenden Varianten und der empirischen Teststärke. Betrachtung des quantitativen Phänotyps ohne Kovariablen. Gen-weise Gruppierung der Varianten mit einer Frequenz des seltenen Allels $\leq 0{,}05$. 98
3.11 Quantil-Quantil-Plots der p-Wert-Verteilung. Betrachtung des quantitativen Phänotyps mit Kovariablen. Gen-weise Gruppierung der Varianten mit einer Frequenz des seltenen Allels $\leq 0{,}01$ und $\leq 0{,}05$. 101
3.12 Streudiagramme der assoziierten Regionen von Interesse bzgl. des Anteils der verursachenden Varianten und der empirischen Teststärke. Betrachtung des quantitativen Phänotyps mit Kovariablen. Gen-weise Gruppierung der Varianten mit einer Frequenz des seltenen Allels $\leq 0{,}01$ und $\leq 0{,}05$. 103

Tabellenverzeichnis

2.1 Eigenschaften der Gruppierungsmethoden 41
2.2 2×2 - Kontingenztabelle: Präsenz bzw. Abwesenheit mindestens eines seltenen Allels in der Region von Interesse bei Fällen und Kontrollen 43
2.3 2×2-Kontingenztabelle: Anzahl der Risiko- und Wildtyp-Allele bei Fällen und Kontrollen bzgl. einer Region von Interesse. 53

3.1 Zusammenfassung der Filterkriterien und der entsprechenden Regionen von Interesse in den Simulationsdaten. 77
3.2 Anzahl assoziierter und nicht-assoziierter Regionen von Interesse der Filter-Szenarien in den Simulationsdaten 82
3.3 Teststärke und Fehler 1. Art bei Gen-weiser Gruppierung aller bzw. ausschließlich nicht-synonymer Varianten mit einer Frequenz des seltenen Allels $\leq$ 0,01 und $\leq$ 0,05. Fall-Kontroll-Status ohne Kovariablen. 87
3.4 Teststärke und Fehler 1. Art bei Gen-weiser Gruppierung aller bzw. ausschließlich nicht-synonymer Varianten mit einer Frequenz des seltenen Allels $\leq$ 0,01 oder $\leq$ 0,05. Fall-Kontroll-Status mit Kovariablen. 91
3.5 Teststärke und Fehler 1. Art bei Gen-weiser Gruppierung aller bzw. ausschließlich nicht-synonymer Varianten mit einer Frequenz des seltenen Allels $\leq$ 0,01 und $\leq$ 0,05. Quantitativer Phänotyp ohne Kovariablen. 99
3.6 Teststärke und Fehler 1. Art bei Gen-weiser Gruppierung aller bzw. ausschließlich nicht-synonymer Varianten mit einer Frequenz des seltenen Allels $\leq$ 0,01 und $\leq$ 0,05. Quantitativer Phänotyp mit Kovariablen. 100

4.1 Anzahl von Varianten für die Regionen von Interesse des Realdatensatzes bei Wahl des quantitativen Phänotyps bzgl. Frequenz des seltenen Allels von $\leq$ 0,05 und $\leq$ 0,01. 108

4.2 Anzahl von Varianten für die Regionen von Interesse des Realdatensatzes bei Wahl des dichotomen Phänotyps bzgl. Frequenz des seltenen Allels $\leq 0{,}05$ und $\leq 0{,}01$. 108
4.3 p-Werte bzgl. dreier Regionen von Interesse des Realdatensatzes in den einzelnen Gruppierungsmethoden 110

1 Einleitung

1.1 Hintergrund

Die Desoxyribonukleinsäure (DNS, engl. DNA) ist Träger der genetischen Erbinformation jedes Menschen. Die DNA kodiert einen Großteil der Informationen, die den Aufbau und die Struktur des Organismus bestimmen. Veränderungen in der DNA können die Ursache von krankhaften körperlichen Veränderungen sein. Um Krankheiten frühzeitig zu erkennen bzw. wirksam behandeln zu können, besteht eine zentrale Aufgabe der medizinischen und epidemiologischen Forschung darin, Zusammenhänge zwischen Genveränderungen in der DNA und auftretenden Krankheitsbildern zu erkennen und dafür sowohl entsprechende diagnostische als auch statistische Methoden zu entwickeln (Campbell und Reece 2000).

Die DNA ist ein Polymer, das sich aus Nukleotiden zusammensetzt. Jedes dieser Nukleotide besteht aus einem Phosphat-Rest, dem Zucker Desoxyribose und einer der vier organischen Basen *Adenin*, *Thymin*, *Guanin* und *Cytosin*. Die Form der DNA ist die einer schraubenförmigen Doppelhelix, in der sich immer zwei bestimmte Basen gegenüberliegen, entweder Adenin und Thymin oder Guanin und Cytosin. Durch die unterschiedliche Abfolge der Nukleotide in der DNA ergibt sich für jeden Menschen eine eindeutige Struktur jedes DNA-Strangs. Gene sind Abschnitte der DNA, die den Aufbau von Proteinen oder Molekülen bestimmen. Die Basen-Sequenz innerhalb eines Gens kodiert dabei spezifische Informationen zur Proteinsynthese (Campbell und Reece 2000; Pearson 2006). Die Kodierung erfolgt dabei über Sequenzen, die aus einer Folge von Nukleotidtriplets bestehen. Ein Nukleotidtriplet besteht jeweils aus drei Basen und kodiert eine Aminosäure. Folgen von Aminosäuren werden wiederum für die Proteinsynthese transkribiert, d.h. umgeschrieben. Die gesamte

DNA des Menschen verteilt sich auf 23 Chromosomenpaare. Davon sind 22 dieser Paare homolog, d.h. die Chromosomen stimmen im grundlegenden Aufbau überein. Das „Geschlechtschromosom“ hingegen ist nur bei Frauen homolog, bei Männern hingegen heterolog (Watson und Crick 1953).

Die durch die Gene kodierte Proteinsynthese bestimmt dabei maßgeblich das Erscheinungsbild des einzelnen Menschen. Proteine sind u.a. für den Aufbau eines Großteils des menschlichen Körpers, eine Reihe von Stoffwechselprozessen oder der Abwehr von Infektionen verantwortlich. So beeinflussen sie die Ausprägung verschiedener innerer und äußerer Körpermerkmale, die sogenannten *Phänotypen*. Phänotypen können verschieden skaliert sein, man unterscheidet dabei *qualitative* und *quantitative* Phänotypen. Qualitative Phänotypen lassen sich in diskrete Kategorien einteilen, die einander ausschließen. Ein Spezialfall stellt hierbei der *dichotome* (oft auch als *binär* bezeichnete) Phänotyp dar, der genau zwei Ausprägungen haben kann, wie der *Fall-Kontroll-Status*, der Teilnehmer einer Studie in „krank“ und „gesund“ bzw. „1“ und „0“ einteilt. Quantitative Phänotypen hingegen haben stetige, messbare Ausprägungen wie das Körpergewicht oder der systolische Blutdruck (Ziegler und König 2010).

Kommt es zu einer Veränderung in der Basensequenz eines Gens spricht man im Allgemeinen von einer Mutation (Murken 2006). Eine Mutation liegt vor bei einem Austausch einzelner Basen durch andere oder dem Einbau bzw. Verlust einer Sequenz von Basen. Dies kann höchst unterschiedliche Konsequenzen haben. So bleibt ein Einzelbasenaustausch häufig folgenlos, wenn die Basen synonym ausgetauscht werden und damit das Nukleotidtriplet dieselbe Aminosäure kodiert, wie die ursprüngliche Sequenz. Andererseits kann durch eine andere Base auch eine andere Aminosäure kodiert oder der Ableseprozess für ein Gen abgebrochen werden. Je nach Funktion des Proteins können die Konsequenzen für den Phänotyp kaum messbar sein oder auch schwere Funktionsstörungen zur Folge haben (Ziegler und König 2010).

Viele Krankheiten lassen sich mit einer Veränderung der DNA in Zusammenhang bringen. Als Indiz für eine genetische Ursache einer Erkrankung gilt beispielsweise ein gehäuftes Auftreten von Krankheitsfällen unter verwandten Personen. Um die relevanten Genorte, auch Loci genannt, identifizieren zu können, wurden in Kopplungsanalysen die Vererbungsmuster untersucht (Ziegler und König 2010). In einem

wegweisenden Artikel zeigten Risch und Merikangas (1996), dass die Teststärke, um genetische Effekte moderater Stärke zu finden, in Assoziationsstudien mit unabhängigen, also nicht-verwandten, Personen größer ist, als in Kopplungsanalysen. Durch große technische Fortschritte in der Microarray-Technologie (Katsanis und Katsanis 2013) ist es seit ca. 10 Jahren möglich, genomweite Assoziationsstudien (engl. Genome-wide Association Studies, GWAS) durchzuführen. Diese basieren auf der „Common variant – common disease (CVCD)"-Theorie die besagt, dass innerhalb einer Population häufig vorkommende komplexe Krankheiten u.a. von einer oder mehreren häufigen Mutationen, insbesondere Einzelbasenaustauschen (engl. Single Nucleotide Polymorphism, SNP), verursacht werden (Manolio 2010). Eine Vielzahl von GWAS haben die CVCD – Hypothese seitdem bestätigt, u.a. Hindorff et al. (2009), Manolio et al. (2009), Seng und Seng (2008) und Welter et al. (2014).

Für viele der untersuchten Krankheiten konnte der Ursprung der Gesamt-Variabilität jedoch nicht vollständig erklärt werden. Daraus entstand der Begriff der „verschwundenen" Erblichkeit (Maher 2008; Manolio et al. 2009) und es kam die Frage auf, wodurch die übrige Variabilität, die also weder durch SNPs noch durch Umwelteinflüsse erklärt werden kann, verursacht wird. Eine Antwort kann die „Rare variant – common disease (RVCD)"–Theorie geben (Maher 2008). Sie basiert auf der Hypothese, dass eine häufig auftretende Krankheit durch eine Gruppe von *seltenen* Mutationen bzw. Varianten verursacht wird.

Die Überprüfung der RVCD – Hypothese stellte bis vor wenigen Jahren sowohl aus technischer als auch aus statistischer Sicht eine große Herausforderung dar. So war es kaum möglich, Mutationen mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) von weniger als 5% zuverlässig und kosteneffizient zu detektieren. Statistische Probleme ergeben sich aus der Tatsache, dass die Teststärke für Varianten mit geringer Häufigkeit (unter 5%) zu klein ist. Daher können gängige Assoziationstests wie der χ^2-Test nicht angewendet werden (Burkett und Greenwood 2013).

Das im Jahr 1990 gegründete Humangenomprojekt hatte neben der vollständigen Entschlüsselung des menschlichen Genoms mittels Sequenzierung auch die Entwicklung neuer Sequenzierungstechnologien und die entsprechende Datenanalyse zum Ziel (U.S. Department of Energy und National Institutes of Health 2014). Mit der Fertig-

stellung der vollständigen Sequenz der menschlichen DNA wurde deutlich, dass die technologischen und wissenschaftlichen Möglichkeiten noch viel größer waren, als bis dahin angenommen. Jedoch wurde auch klar, dass wesentlich kostengünstigere und schnellere Technologien notwendig waren, um größere Stichproben zeitnah sequenzieren und analysieren zu können. Aus diesem Grund wurde bald darauf vom National Human Genome Research Institute ein weiteres Projekt gegründet, dass insbesondere die Senkung der Kosten zur Sequenzierung des menschlichen Genoms auf 1000 US$ je Person erreichen wollte (Dijk et al. 2014). Dies war der Auftakt für die Entwicklung einer Reihe von Sequenzierungstechnologien der nächsten Generation (engl. Next Generation Sequencing, NGS) (Metzker 2010). Zu den ersten veröffentlichten Techniken gehörten die Pyrosequenzierungsmethode von Life Science (heute Roche) (Margulies et al. 2005) und die Solexa/Illumina Sequenzierungsplattform (Bennett 2004; Bentley et al. 2008) sowie die Technologie Sequencing by Oligo Ligation Detection (SOLiD) (Valouev et al. 2008). Im Vergleich zur ersten Generation der Sequenzierungstechnik, der Sanger-Sequenzierung, mussten DNA-Fragmente nun nicht mehr mit Hilfe bakterieller Klonierung vervielfältigt werden, sondern Fragment-Bibliotheken wurden Zell-frei angelegt. Anstelle von einigen Hundert war es nun möglich, bis zu mehrere Millionen paralleler Sequenzierungsreaktionen durchzuführen. Außerdem verkürzte sich die Weiterverarbeitung des Sequenzierungsergebnisses deutlich im Vergleich zur Sanger-Sequenzierung. Eine detaillierte Beschreibung dieser Techniken findet sich in (Dijk et al. 2014; Liu et al. 2012; Metzker 2010). Im vergangenen Jahr veröffentlichte Illumina *HiSeq X Ten*, eine aus zehn HiSeq X Sequenzierungsmaschinen bestehende Anlage, die eine Sequenzierung der gesamten DNA eines Individuums für 1000 US$ erstellt. Im Vergleich zu den Kosten für die Sequenzierung im Humangenomprojekt, die bei 10 Mio. US$ lag, bedeutet dies eine Reduktion um das 10.000-fache. Jedoch sind die Anschaffungskosten für die Maschinen der HiSeq X Ten mit ebenfalls 10 Mio. US$ sehr hoch (Dijk et al. 2014), so dass in naher Zukunft ihre Nutzung nur für wenige Forschungseinrichtungen finanziell möglich sein wird.

Die enorme Weiterentwicklung der NGS-Technologien in den letzten Jahren ermöglicht die Untersuchung vieler Fragestellungen, die zuvor unbeantwortet bleiben mussten (Dolled-Filhart et al. 2013). Dies rückte auch die RVCD-Theorie erneut in den Fokus, allerdings war noch immer eine Weiterentwicklung der statistischen Testmethoden notwendig, da die bis dahin zur Verfügung stehenden Tests aufgrund der geringen Teststärke keine zuverlässigen Ergebnisse lieferten. Eine Vielzahl neuer Ansätze zur

Analyse von seltenen Varianten wurde vorgeschlagen (Basu und Pan 2011). Ein Teil dieser Ansätze wird in der vorliegenden Arbeit untersucht.

Ein Ansatz zur Untersuchung der RVCD-Theorie besteht darin, die Genotyp-Informationen einer Gruppe von Varianten, die einer biologisch bzw. funktionell zusammenhängenden Einheit, i.d.R. ein Gen oder eine Gengruppe, angehören, in einer Teststatistik zusammenzufassen und gemeinsam zu untersuchen. Die dadurch gebildete Gruppe wird Region von Interesse (engl. Region of Interest, ROI) genannt (Basu und Pan 2011).

Basierend auf diesem Gruppierungsansatz ist eine Reihe von statistischen Methoden entstanden, die sich unter dem Begriff *Gruppierungsmethoden* (engl. *Collapsing Methods*) zusammenfassen lassen (Li und Leal 2008). Durch die Bildung einer ROI wird nicht nur die effektive MAF, d.h. die kumulierte MAF der Varianten einer ROI, sondern möglicherweise auch der zu messende Effekt der gebildeten Gruppe erhöht. Im besten Fall kann damit ein vorhandenes Assoziationssignal verstärkt und detektiert werden.

1.2 Bisherige Vergleichsarbeiten zu Gruppierungsmethoden

Zu Beginn dieses Jahrhunderts wurden in den Arbeiten von Cohen et al. (2004) und Fitze et al. (2002) erstmals Gruppierungen seltener Varianten untersucht. Daraufhin folgte die Veröffentlichung einer Vielzahl von weiteren Gruppierungsmethoden, u.a. Li und Leal (2008), Madsen und Browning (2009), Morgenthaler und Thilly (2007) und Morris und Zeggini (2010). Viele dieser Gruppierungsansätze wurden in kleineren Vergleichsstudien vorgestellt und mit einigen der bis dahin publizierten Ansätze verglichen (Basu und Pan 2011; Lee, Emond et al. 2012; Liu und Leal 2010; Luo et al. 2011; Wu et al. 2011a; Zawistowski et al. 2010). Die den Methoden zugrundeliegenden Ideen basierten in der Regel auf unterschiedlichen Annahmen bzgl. der genetischen Struktur der seltenen Varianten und deren Assoziation zu einer häufigen Krankheit, was zu verschiedenen Simulationsansätzen führte (Stitziel et al. 2011). Zudem wurde häufig auf die Betrachtung des Einflusses von Kovariablen in den zugehörigen

Simulationsstudien verzichtet. Aus diesem Grund erlauben die Ergebnisse dieser Vergleichsarbeiten keine unabhängige und allgemeingültige Leistungsbewertung der einzelnen Methoden.

Eine der ersten methodenunabhängigen Vergleichsstudien ist die Übersichtsarbeit von Bansal et al. (2010). Neben diversen Gruppierungsmethoden wurden auch Probleme bei der Untersuchung seltener Varianten diskutiert, wie z.B. die zugrundeliegende Sequenzierungstechnik, Genotypisierungsfehler, Stratifikation und die geeignete Wahl von Imputationstechniken. Die Arbeit gibt einen umfassenden Überblick über die bis dato existierenden Gruppierungsmethoden, bietet aber keinen allgemeingültigen Leistungsvergleich, da lediglich die Angaben aus den Originalarbeiten herangezogen wurden.

Ladouceur et al. (2012) untersuchten in einer Stichprobe von ca. 2000 Individuen den Einfluss verschiedener Aspekte auf die Teststärke von fünf Gruppierungsmethoden. In der Studie wurden u.a. der Einfluss der Effektstärken einzelner Varianten, der Anteil verursachender Varianten in sieben unterschiedlichen Regionen von Interesse und die Art des betrachteten Phänotyps untersucht. Die Autoren beschränkten sich dabei auf fünf Gruppierungsmethoden. Auf eine Untersuchung der Einhaltung des Fehlers 1. Art wurde gänzlich verzichtet.

In der Arbeit von Burkett und Greenwood (2013) wird ein allgemeiner Überblick über verschiedene Ansätze zur Untersuchung von seltenen Varianten in Bezug auf häufige Krankheiten gegeben. Darunter sind auch einige der hier diskutierten Methoden, bei der die Gruppierung der Varianten die Basis der Untersuchung bildet. Allerdings verzichten die Autoren auf einen Leistungsvergleich dieser Ansätze.

Im Fokus der Übersichtsarbeit von Lee et al. (2014), stehen, ähnlich wie bei Bansal et al. (2010), sowohl technische als auch statistische Probleme in der Analyse von seltenen Varianten. Neben den Vorschlägen zu möglichen Studiendesigns werden auch die verschiedenen Prinzipien einer Vielzahl statistischer Methoden zur Analyse seltener Varianten betrachtet, darunter einige der hier betrachteten Gruppierungsmethoden. Jedoch wird keine unabhängige Simulationsstudie durchgeführt. Des Weiteren gehen Lee et al. (2014) detailliert auf die Planung und das Vorgehen bei

einer Assoziationsstudie mit seltenen Varianten ein, geben aber kaum Hinweise zur Definition einer geeigneten Region von Interesse.

Derkach et al. (2014) definieren in ihrer Arbeit zwei Klassen von Teststatistiken bzgl. der Gruppierungsmethoden, auf die in Abschnitt 2.2.5 näher eingegangen wird. Zentraler Punkt der Arbeit ist eine theoretische Untersuchung der Teststärke dieser beiden Klassen bzgl. verschiedener simulierter genetischer Strukturen der Regionen von Interesse, unterschiedlicher Stichprobenzahl und verschieden skalierter Phänotypen. Wie auch in früheren Vergleichsstudien werden keine Kovariablen betrachtet.

In der erst vor kurzem erschienenen Arbeit von Moutsianas et al. (2015) wird zum einen eine Software zur Erstellung unabhängiger Simulationsdaten vorgestellt. Zum anderen wird ein Vergleich eines Teils der hier betrachteten Gruppierungsmethoden auf Basis diverser Kriterien wie Stichprobengröße, Präsenz neutraler Varianten und vorhandener Effektstruktur durchgeführt. Der Fokus der Vergleichsuntersuchungen liegt in der Betrachtung des binären Phänotyps des Fall-Kontroll-Status ohne eine Betrachtung von Kovariablen.

1.3 Ziele der Arbeit

Im Folgenden soll ein ausführlicher Leistungsvergleich von 15 verschiedenen Gruppierungsmethoden gegeben werden. Zu diesem Zweck werden für einen Simulationsdatensatz der Fehler 1. Art und die Teststärke für mehrere Untersuchungsszenarien ermittelt und hinsichtlich möglicher Einflussfaktoren dieser Maße miteinander verglichen (Almasy et al. 2011). Schließlich werden die 15 Gruppierungsmethoden auf einen Realdatensatz angewendet, um die aus den Ergebnissen bzgl. des Simulationsdatensatzes gezogenen Schlüsse überprüfen zu können. Hierbei wird die Assoziation des Methotrexat-Abbaus bei an akuter lymphatischer Leukämie erkrankten Kindern mit dem Gen *SLCO1B1* betrachtet (Ramsey et al. 2012; Treviño et al. 2009).

Die in Abschnitt 1.2 genannten Arbeiten, insbesondere die Vergleichsstudien (Bansal et al. 2010; Burkett und Greenwood 2013; Derkach et al. 2014; Ladouceur et al.

2012; Lee et al. 2014; Moutsianas et al. 2015), bilden eine wichtige Grundlage für die vorliegende Untersuchung. Insgesamt wird im Folgenden aber ein breiteres Spektrum an Gruppierungsmethoden untersucht und verglichen. Zudem werden neue Aspekte in Betracht gezogen: Zum einen die Untersuchung von quantitativen und binären Phänotypen, zum anderen werden zusätzlich Kovariablen einbezogen. Zur statistischen Validierung werden dabei bei den Permutations-basierten Ansätzen mehr Permutationen als sonst üblich verwendet. Zum Zwecke der Anwendung der Gruppierungsmethoden in Assoziationsstudien wird eine detaillierte Anleitung zur Bildung einer Region von Interesse gegeben. Die einzelnen Gruppierungsmethoden werden erklärt, algorithmisch beschrieben und in ihren Eigenschaften betrachtet. Das Prinzip der einzelnen Gruppierungsmethoden wird dabei sowohl im biologischen als auch im statistischen Kontext betrachtet. Neben der Überprüfung der Erkenntnisse der vorgenannten Vergleichsstudien ist die zentrale Errungenschaft der vorliegenden Arbeit eine unabhängige Leistungsbewertung der hier betrachteten Gruppierungsmethoden in Bezug auf einen Simulations- und einen Realdatensatz unter Einbeziehung verschiedener Aspekte. Dies ist insbesondere für Anwender von Gruppierungsmethoden von Bedeutung.

Die Wahl der 15 Gruppierungsmethoden basiert auf verschiedenen Kriterien. Zum einen wurden die frühen Gruppierungsmethoden untersucht. Dies ermöglicht nicht nur den Vergleich der Ergebnisse mit vorhergehenden Vergleichsstudien, sondern stellt auch sicher, dass die am häufigsten verwendeten Methoden betrachtet werden. Zum anderen wurde darauf geachtet, Methoden mit möglichst vielen verschiedenen Gruppierungskonzepten zu verwenden. Dies ermöglicht einen umfassenden Vergleich hinsichtlich der verschiedenen Eigenschaften. Aufgrund der rasanten Entwicklung und der vielen Veröffentlichungen von neuen Gruppierungsansätzen ist ein Vergleich aller Methoden jedoch nicht möglich. So gibt es beispielsweise bereits weit über 50 statistische Tests zur Analyse seltener Varianten (Auer und Lettre 2015; Basu und Pan 2011; Lee et al. 2014; Moutsianas et al. 2015), von denen im Folgenden insbesondere ein Teil der Gruppierungsmethoden, bei denen die Gruppierung mehrerer seltener Varianten im Fokus steht, betrachtet werden sollen. Auf die potentielle Verwandtschaft der einzelnen Ansätze zu anderen hier nicht untersuchten Gruppierungsmethoden wird in den jeweiligen Methodenbeschreibungen hingewiesen. Dennoch kann diese Arbeit keinen Anspruch auf Vollständigkeit in der Untersuchung aller existierenden Gruppierungskonzepte erheben. Neue Methoden sowie Möglichkeiten zur Weiterentwicklung

mit Bezug auf Erkenntnisse, die aus dieser Studie gezogen werden konnten, werden in einem Ausblick am Ende der Arbeit diskutiert.

Mit Bezug auf die verschiedenen Gruppierungskonzepte ist ein weiterer Schwerpunkt der Arbeit die Darlegung und Untersuchung definierender Charakteristika von Gruppierungsmethoden im Hinblick auf ihre Relevanz in der praktischen Anwendung. Die Gruppierungsmethoden unterscheiden sich in einer Reihe von Eigenschaften, wie der Interpretation der Gruppe von Varianten, der Verwendung von Gewichten in den zugehörigen Teststatistiken, der Schätzung der entsprechenden p-Werte oder den möglichen Anwendungssituationen (Lee et al. 2014). Ein umfassender Vergleich hinsichtlich dieser Eigenschaften sowie der statistischen Validität unter Berücksichtigung verschieden skalierter Phänotypen und Kovariablen wurde bisher nicht durchgeführt, insbesondere nicht für den bisher noch realistischen Fall von kleinen Stichprobengrößen. Im Zuge dessen wird auch eine Schritt-für-Schritt-Anleitung für die Implementation aller Methoden gegeben. Für einige Algorithmen existieren aufgrund uneindeutiger Methodenbeschreibungen unterschiedliche Software-Implementierungen, was zu Unterschieden bei den berechneten Ergebnissen führen kann.

Ein weiteres Ziel der Arbeit liegt in der Erarbeitung einer generellen Vorgehensweise zur Bildung einer ROI sowie auf der Überprüfung geeigneter Filterkriterien zur Bestimmung einer ROI. Grund dafür sind verschiedene Probleme, die bei der Gruppierung der Varianten zu ROIs auftreten können: Zum einen ist die konkrete Zusammensetzung einer solchen Gruppe ohne Vorwissen zunächst unklar, was in der Praxis eine Vielzahl von zu untersuchenden Regionen und somit einen enormen Rechenaufwand bedeuten würde. In der Regel wird eine ROI auf Basis eines Gens oder einer Gruppe von Genen definiert. Dabei kann es zum anderen passieren, dass Varianten mit beiden Effektrichtungen, positiv und negativ, oder auch neutrale Varianten in die Gruppe integriert werden. Dass sich neutrale Varianten in der Gruppe befinden, passiert in der Mehrzahl der Fälle. Dies kann eine Verwässerung der vorhandenen Einzeleffekte zur Folge haben. Die Folge für die Präsenz von Varianten mit positiven und negativen Effekten kann eine Auslöschung der einzelnen Effekte sein. Dies kann wiederum zu Schwierigkeiten bei der Detektion des Effekts der Region führen (Manolio et al. 2009; Zhang et al. 2011). Um die RVCD-Hypothese zu prüfen, werden für die meisten Gruppierungsmethoden alle seltenen Varianten der funktionellen Einheit zusammengefasst. Neben der Wahl der MAF der Varianten ist die Anwendung weite-

rer Filterkriterien, wie z.B. der Lage innerhalb eines Gens oder der Funktionalität der Varianten zur Bildung einer ROI denkbar (Price et al. 2010). Diesbezüglich erfolgt eine Zusammenstellung geeigneter Software-Programme zur Hinzufügung von Filterkriterien sowie eine Überprüfung der Eignung der Filterkriterien.

1.4 Aufbau der Arbeit

Zu Beginn von Kapitel 2 werden Datenbanken und Software diskutiert, die der Gewinnung von Informationen zur Lokalisierung von Varianten im Genom und weiterer Eigenschaften der Varianten dienlich sein können. Anschließend wird eine Anleitung zur Bestimmung einer Region von Interesse (ROI) gegeben. Dabei wird von der grundlegenden Definition der *einfachen* Region von Interesse ausgegangen. Anschließend werden Filterkriterien diskutiert, um eine *gefilterte* Region von Interesse zu bestimmen.

In Abschnitt 2.2 werden strukturelle Charakteristika der zugehörigen Teststatistiken und Eigenschaften der Gruppierungsmethoden vorgestellt und bzgl. ihrer möglichen Relevanz in der praktischen Anwendung diskutiert. Der darauffolgende Abschnitt 2.3 beinhaltet eine detaillierte Darlegung der entsprechenden Algorithmen und Eigenschaften der 15 verwendeten Gruppierungsansätze in chronologischer Reihenfolge.

Kapitel 3 umfasst die zentralen Ergebnisse der Arbeit. Alle 15 Gruppierungsmethoden werden anhand des Simulationsdatensatzes des *Genetic Analysis Workshop 17* (Almasy et al. 2011) auf ihre statistische Validität hin überprüft. Hierfür werden auf Basis von 3205 Genen und verschiedenen Filterkriterien ROIs gebildet. Die daraus entstehenden Szenarien werden jeweils mit quantitativen und qualitativen Phänotypen mit und ohne Kovariablen in allen geeigneten Gruppierungsmethoden untersucht.

Alle Gruppierungsmethoden werden in Kapitel 4 auf einen Realdatensatz angewendet. Dabei werden die Unterschiede in der Signifikanz hinsichtlich drei verschiedener ROIs bzgl. der Assoziation mit einem quantitativen Phänotyp untersucht.

In Kapitel 5 erfolgt eine ausführliche Diskussion der Ergebnisse und der sich ergebenden Konsequenzen für die praktische Anwendung. Außerdem wird ein Ausblick zu weiteren Analysemethoden für seltene Varianten sowie zukünftigen Untersuchungsfeldern gegeben.

2 Material und Methoden

2.1 Datenvorverarbeitung: Bildung einer Region von Interesse

In der RVCD-Theorie wird angenommen, dass eine Gruppe von seltenen Varianten gemeinsam eine häufig auftretende Krankheit verursacht (Maher 2008). Für die Überprüfung dieser Theorie mit Hilfe einer Gruppierungsmethode ist die konkrete Definition einer Gruppe von Varianten erforderlich. Wie sich eine solche Gruppe von seltenen Varianten genau zusammensetzt, ist bis auf, dass sie in der Regel auf einem Gen basiert (Derkach et al. 2014; Li und Leal 2008), unbekannt. In diesem Abschnitt soll ein möglicher allgemeiner Ablauf zur Definition einer ROI beschrieben werden. Eine zusammenfassende Darstellung der Vorgehensweise findet sich in Abbildung 2.1. In den Abschnitten 2.1.3 und 2.1.4 wird auf die Bildung einer einfachen bzw. gefilterten ROI im Detail eingegangen.

Im Folgenden wird davon ausgegangen, dass zu einer Stichprobe Sequenzdaten vorliegen, die durch NGS-Technologien (Metzker 2010) erstellt und anschließend genotypisiert wurden. Des Weiteren wird angenommen, dass die so erhaltenen Daten bereits einer Qualitätskontrolle unterzogen wurden und dass es keine fehlenden Werte bzgl. der Genotypdaten in der Stichprobe gibt.

Die aus der Sequenzierung und anschließenden Genotypisierung erhaltenen Genotypdaten werden dann mit Hilfe von sogenannten *Annotationsdatenbanken* (s. Abschnitt 2.1.1) mit Informationen versehen, die für die Lokalisierung der Varianten im Genom notwendig sind oder weitere wichtige Eigenschaften der Varianten beinhalten. Welche Informationen hierbei im Einzelnen annotiert werden, ist je nach

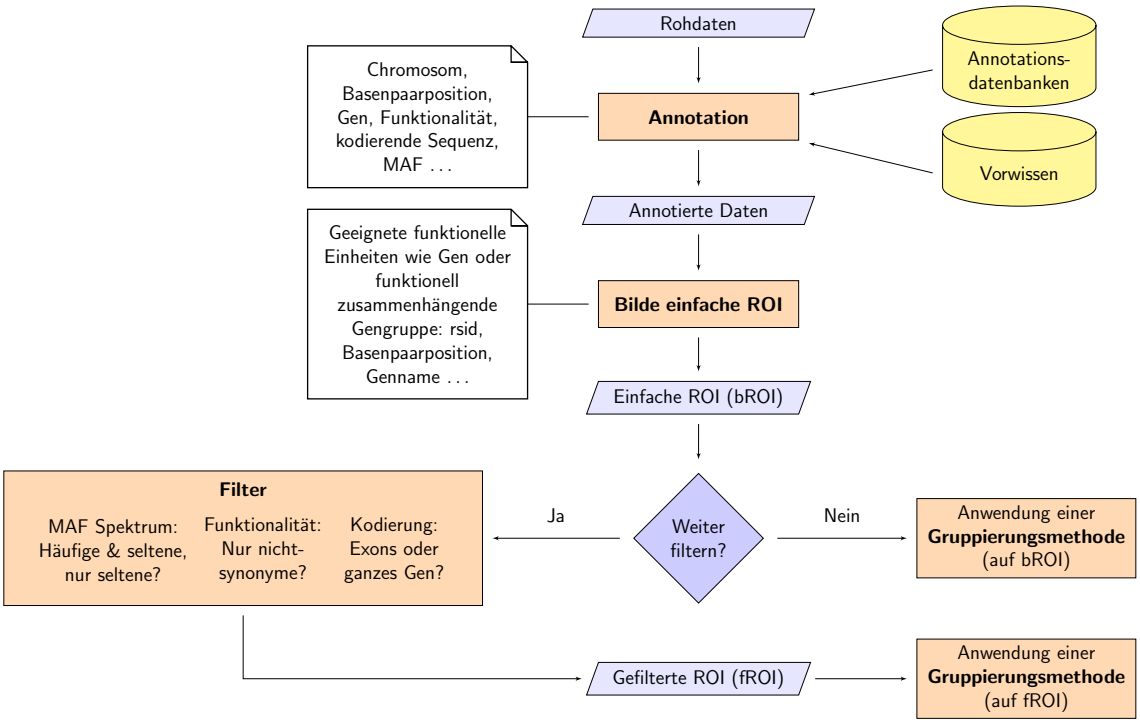


Abbildung 2.1: Flussdiagramm zur Bildung einer Region von Interesse

Fragestellung manuell auszuwählen. Notwendige Informationen sind u.a. die Basenpaarpositionen und die Chromosomennummern. Da ein Gen häufig als funktionelle Einheit als Basis zur Definition einer ROI verwendet wird, ist es sinnvoll, evtl. den entsprechenden Gen-Namen ebenfalls zu annotieren. Anhand dieser drei Informationen und der ursprünglichen Genotypdaten kann mit Hilfe von möglicherweise vorhandenem Vorwissen eine sog. *einfache Region von Interesse* (engl. basic Region of Interest, kurz bROI) definiert werden. Das Vorwissen kann eine vorangegangene Studie sein, bei der bereits eine Assoziation zwischen einzelnen SNPs einer Region oder eines Gens mit einem Phänotyp gefunden wurde. Dann werden beispielsweise die Genotypdaten der Stichprobe nur auf diese Region bzw. dieses Gen eingeschränkt. Sobald eine einfache ROI definiert wurde, ist man in der Lage, sie mit Hilfe einer Gruppierungsmethode zu untersuchen.

Die einfache ROI kann aber noch genauer spezifiziert und damit verkleinert werden, indem man weitere annotierte Informationen als Filterkriterien verwendet und auf deren Basis eine sogenannte *gefilterte Region von Interesse* (engl. filtered Region of Interest, kurz fROI) bestimmt. Dies kann sinnvoll sein, wenn einzelne Gruppierungsmethoden bestimmte Schwächen aufweisen, deren Auswirkungen durch die

Filterung eingeschränkt werden können (vgl. Abschnitt 2.2.2)(Lee, Emond et al. 2012; Zhang et al. 2011). Mögliche Filterkriterien sind z.B. die Frequenz des seltenen Allels (MAF), die Protein-kodierenden Regionen eines Gens (Exon) oder die Information, ob die Variante die Aminosäure synonym (S) oder nicht-synonym (NS) kodiert. Hat man die Filterkriterien festgelegt, kann die bROI entsprechend gefiltert werden und man erhält eine fROI, auf die nun ebenfalls eine Gruppierungsmethode angewendet werden kann.

2.1.1 Annotationsdatenbanken

Zur Identifizierung kausaler Varianten bzw. zur Eingrenzung der Menge potentiell kausaler Varianten ist eine geeignete *Annotation* unerlässlich. Annotation meint hierbei die Aggregierung von Informationen bzgl. der innerhalb einer Studie zur Verfügung stehenden Sequenzdaten bzw. Varianten, die zum einen die eindeutige Lage der Varianten innerhalb des Genoms bestimmen und zum anderen für den betrachteten biologischen Prozess von Bedeutung sind. Auf Basis solcher zusätzlichen Informationen lässt sich die zu untersuchende Menge von Varianten nach Anwendung von geeigneten Filterkriterien (Abschnitt 2.1.2) erheblich reduzieren. Allgemeine Informationen zur Lokalisierung der Varianten wie Chromosomennummer, Basenpaarposition oder zusätzlich das entsprechende Gen werden häufig in speziellen Annotationsdateien für die verwendete Sequenzierungstechnologie von den Herstellern mitgeliefert. Darüber hinausgehende Informationen wie z.B. die Angabe von Varianten-Identifizierungsnummern in verschiedenen Datenbanken sowie Informationen über den zugehörigen funktionellen Effekt können in speziellen Datenbanken abgefragt werden. Zu den bekanntesten Datenbanken zählen hierbei *Ensembl*, *dbSNP* und die *UCSC*-Datenbank (Cunningham et al. 2015; Rosenbloom et al. 2015; Sherry et al. 2001). In ihnen sind Informationen über bekannte häufige und seltene Varianten gespeichert, die bzgl. eines Referenzgenoms und bei neuen Studienergebnissen aktualisiert werden.

Ein universelles und stets aktuell gehaltenes Annotations-Abfragewerkzeug ist die frei zugänglichen Software *Annotate Variation*, kurz: *ANNOVAR* (Wang et al. 2010). Sie ermöglicht eine simultane Datenabfrage aus etlichen öffentlich zugänglichen Datenbanken, unabhängig von der verwendeten Sequenzierungstechnologie und je

nach Bedarf für einzelne oder eine Liste von Varianten. *ANNOVAR* bietet außerdem die Möglichkeit zur Annotation von Genom-Daten verschiedener Spezies, sowie von Insertionen und Deletionen, sog. Indels. Dabei handelt es sich um einen einzelnen oder Basen-weisen Einbau bzw. Verlust von bis zu 50 Einzelbasen im Vergleich zu einem Referenzgenom. Außerdem ist *ANNOVAR* in der Lage, gängige Bewertungsmaße hinsichtlich der Schädlichkeit nicht-synonymer Varianten anzugeben, (z.B. *SIFT*-(Ng und Henikoff 2003) oder *PolyPhen2*-Score (Gorlov et al. 2008)), die in Abschnitt 2.1.2.2 näher erläutert werden. Des Weiteren nutzt *ANNOVAR* Informationen aus dem 1000 Genome Projekt (1000 Genomes Project Consortium et al. 2010), welches insbesondere relevant für seltene Varianten ist, die in der Regel nicht in den gängigen Datenbanken enthalten sind. Neben der Gen-basierten Annotation bietet *ANNOVAR* auch die Möglichkeit spezielle Regionen des menschlichen Genoms (konservierte Regionen, Mikro-RNA oder RNA-Strukturen) zu annotieren. Dies ist insbesondere dann interessant wenn sich die meisten Varianten außerhalb der Protein-kodierenden Regionen (den Genen) befinden. Ein Nachteil von *ANNOVAR* ist, dass die zur Annotation verwendeten Informationen lokal gespeichert und somit bei jeder Aktualisierung erneut heruntergeladen werden müssen, was sehr zeitaufwändig sein kann.

In der Arbeit von Dolled-Filhart et al. (2013) werden neben *ANNOVAR* weitere Datenbanken zur Annotation diskutiert. Das *Variant Annotation, Analysis and Search Tool* (*VAAST*) identifiziert mittels probabilistischer Methoden Individuen-basiert, durch seltene und häufige Varianten beschädigte Gene (Yandell et al. 2011). Das *Variant Analysis Tool* (*VAT*) dient insbesondere der Annotation von Funktionsverlust-Varianten, die innerhalb des 1000 Genome Projekts identifiziert wurden (Habegger et al. 2012). Außerdem ist *VAT* in der Lage Indels und sehr seltene Varianten sowohl für nicht-Protein-kodierende (Introns) als auch für Protein-kodierende Abschnitte (Exons) eines Gens zu annotieren. Das *VARIant ANalysis Tool* (*VARIANT*) trägt ähnlich wie *ANNOVAR* Informationen aus verschiedenen Datenbanken wie *dbSNP*, *1000 Genome Projekt*, dem GWAS Katalog, *Online Mendelian Inheritance in Man* (*OMIM*) und *Cosmic*, zusammen (1000 Genomes Project Consortium et al. 2010; Forbes et al. 2015; MacArthur et al. 2012; McKusick-Nathans Institute et al., 2015; Sherry et al. 2001). Der Vorteil von *VARIANT* ist, dass die Informationen initial und nach einer Aktualisierung nicht komplett heruntergeladen und gespeichert werden. Vielmehr werden die benötigten Daten selektiv aus der aktuell gehaltenen, externen

Datenbank über das Internet bezogen. Allerdings kann auch dieser Vorgang sehr zeitintensiv sein. Die *SeattleSeq* Datenbank ist eine der wenigen Datenbanken, die auch Informationen bzgl. neu gefundener Varianten annotieren kann (National Heart, Lung and Blood Institute, 2015).

2.1.2 Filterkriterien

In diesem Abschnitt wird auf die wichtigsten Filterkriterien zur Bildung einer fROI eingegangen und verschiedene Anwendungsmöglichkeiten und Kategorien werden näher erläutert.

2.1.2.1 Minor Allele Frequency (MAF)

Die Frequenz des seltenen Allels stellt für die meisten Gruppierungsmethoden ein essentielles Filterkriterium dar. Dies liegt in der RVCD-Theorie begründet, die nicht nur annimmt, dass eine Gruppe von seltenen Varianten zusammen eine Krankheit verursacht, sondern auch, dass sich durch eine Akkumulation seltener Varianten auch deren Effekte aufsummieren und dann erst eine moderate detektierbare Größe erreicht. Diese Hypothese konnte u.a. in den Arbeiten von Ahituv et al. (2007), Cohen et al. (2004), Fitze et al. (2002) und Nejentsev et al. (2009) validiert werden.

In der Literatur wird die Frequenzschwelle, unterhalb der eine Variante als selten betrachtet werden kann, unterschiedlich angesetzt, die Angaben schwanken zwischen 0,1% und 5% (Bodmer und Bonilla 2008; Lee, Emond et al. 2012; Lin und Tang 2011). Die untere MAF-Grenze erklärt sich aus einer Aussage von Bodmer und Bonilla (2008), wonach die Frequenz schädlicher Varianten in einer Population trotz Selektion höchstens 0,1% beträgt. Als obere MAF-Grenze für seltene Varianten wird im Allgemeinen die untere Grenze für häufige Einzelbasenaustausche angesetzt. Diese liegt theoretisch bei 1%, in der Praxis wird die untere Grenze für Einzelbasenaustausche für GWAS allerdings bei 5% angesetzt, so dass sich auch für bei der Analyse von seltenen Varianten diese Grenze häufig als Obergrenze etabliert hat.

Es kann sinnvoll sein, neben seltenen Varianten zusätzlich auch häufige Varianten in eine Untersuchung einzubeziehen, wie einige Beispiele von Assoziationsstudien zeigen (Brunham et al. 2006; Ramsey et al. 2012; Spirin et al. 2007). In diesen Fällen konnte zusätzlich zur Assoziation eines häufigen Einzelbasenaustauschs mit einem Phänotyp ebenso ein Einfluss von seltenen Varianten der gleichen Region auf die Ausprägung des Phänotyps nachgewiesen werden. Weitere Aspekte bei der Filterung nach der MAF werden in Abschnitt 2.2.3 als Eigenschaft von Gruppierungsmethoden diskutiert.

Im Wesentlichen gibt es drei Möglichkeiten um nach der MAF zu filtern:

Feste Grenze: Falls es eine zuvor festgelegte MAF-Grenze τ_{MAF} gibt, sollten alle Varianten mit einer MAF $> \tau_{\text{MAF}}$ aus der Analyse ausgeschlossen werden. So wird in dem Beispiel in Abbildung 2.3 vorgegangen, wo $\tau_{\text{MAF}} = 5\%$.

Optimale Grenze: Falls es keine feste MAF-Grenze gibt, besteht die Möglichkeit, den Ansatz von Price et al. (2010) zu verwenden und die MAF-Grenze zu suchen, unter der die Signalstärke am größten ist. Die Idee der Autoren basiert auf der Annahme, dass es eine MAF-Grenze gibt, unter welcher alle Varianten der ROI funktionell relevant sind.

Feste Grenze + häufige Varianten: Bei der Betrachtung von seltenen und häufigen Varianten wird ein Verfahren in zwei Schritte unterteilt. Zunächst werden die seltenen Varianten nach einer festen MAF-Grenze τ_{MAF} gruppiert. Zusätzlich werden auch die häufigen Varianten mit einer MAF $> \tau_{\text{MAF}}$ einzeln in die Analyse einbezogen. Es gibt Gruppierungsmethoden wie die von Li und Leal (2008), die diesen Ansatz von vornherein verfolgen (s. Abschnitt 2.3.2). Andere Methoden verwenden sowohl seltene als auch häufige Varianten in einem Schritt zur Berechnung der Teststatistik, geben den seltenen Varianten z.B. auf Basis der MAF ein höheres Gewicht als den häufigen(Lee, Emond et al. 2012; Wu et al. 2011b). Die gemeinsame Untersuchung von seltenen und häufigen Varianten wird speziell in Abschnitt 2.2.3 diskutiert.

2.1.2.2 Funktionalität von Varianten

Im Allgemeinen stellt schon die Gruppierung der Varianten eines Gens eine Gruppierung bzgl. der Funktionalität dar, da ein Gen als eine funktionelle Einheit betrachtet wird, über die, vereinfacht ausgedrückt, einzelne Stoffwechselprozesse eines Individuums gesteuert werden. Allerdings können die funktionellen Einheiten noch wesentlich spezifischer gefasst werden, so dass z.B. nur Varianten gruppiert werden, die einen direkten Einfluss auf die Bildung der entsprechenden Aminosäure oder des zugehörigen Proteins haben:

Pathway: Ein *Pathway* ist eine Gruppe von Genen, die für einen Stoffwechselprozess verantwortlich ist. Um eine Stoffwechselstörung verursachende Variante innerhalb eines bekannten Pathway zu detektieren, kann die Gruppierung aller Varianten des Pathways oder auch nur bestimmter Subgruppen hilfreich sein (Burkett und Greenwood 2013).

Exon (E)/Intron (I): Ein Gen wird in *Intron*- und *Exon*-Abschnitte unterteilt. Die zu einer kontinuierlichen Basenpaar-Folge zusammengefügten Exon-Abschnitte kodieren Aminosäuren, welche wiederum in Proteine transkribiert werden (Campbell und Reece 2000). Daher kann es sinnvoll sein, nur Varianten aus Exon-Abschnitten zu gruppieren. Genomweite Assoziationsstudien haben allerdings gezeigt, dass ca. 88% der schwach mit einem Phänotyp assoziierten Varianten in Intron-Abschnitten zu finden sind (Dolled-Filhart et al. 2013). Je nach Fragestellung kann also die Gruppierung von Varianten ausschließlich aus Intron-Abschnitten oder eben auch eine Gruppierung aller Varianten eines Gens sinnvoll sein.

Synonym (S)/Nicht-Synonym (NS): Varianten aus Exon-Abschnitten können die zugehörige Aminosäure entweder *synonym* oder *nicht-synonym* kodieren. Eine synonyme Kodierung bedeutet, dass die Präsenz der Variante keine Veränderung der Aminosäure zur Folge hat, sondern dieselbe Aminosäure gebildet und somit auch keine Veränderung des entsprechenden Proteins hervorgerufen wird (Ziegler und König 2010). Dem gegenüber steht die nicht-synonyme Kodierung einer Variante bzw. der zugehörigen Aminosäure, welche nicht nur die Bildung einer anderen Aminosäure, sondern auch Stoffwechsel-beeinflussende

Veränderungen wie z.B. der Bindungseigenschaften oder der Polung des kodierten Proteins hervorrufen können (Ramensky et al. 2002).

Schädlich/Unschädlich: Die Folgen einer NS Variante können hinsichtlich der Ausprägung eines bestimmten Phänotyps für ein Individuum *schädlich* oder *unschädlich* sein. Um diese Klassifizierung für eine Variante treffen zu können, werden Bewertungsmaße, die i.d.R. einer Schädlichkeits-Wahrscheinlichkeit entsprechen, mit Hilfe bioinformatischer Software ermittelt. Eine Methode ist z.B. die *Sorting Intolerant from Tolerant* (*SIFT*) Software, die auf der Untersuchung von konservierten genetischen Regionen basiert um potentiell schädliche Mutationen zu finden (Ng und Henikoff 2003). *PolyPhen2* und *MutationTaster* sind Programme, die Methoden des maschinellen Lernens zur Klassifizierung der Varianten und deren Wirkung auf die Proteinbildung verwenden (Adzhubei et al. 2010; Schwarz et al. 2014). Weitere Klassifizierungs-Methoden sind *SNPeff* (Cingolani et al. 2012), *PMUT* (Ferrer-Costa et al. 2004) und *SNPS3D* (Yue et al. 2006). Bewertungsmaße einiger dieser Vorhersage-Programme sind die bereits in den in Abschnitt 2.1.1 vorgestellten Annotationswerkzeugen bzw. -datenbanken integriert, so dass eine gesonderte Berechnung nicht notwendig ist.

Eine mögliche Annotation und Filterung der Varianten einer Stichprobe ist in Abbildung 2.3 dargestellt. Hier wird aus einer einfachen ROI des fiktiven Gens *FMG12* mittels der Kriterien „MAF $\leq 5\%$“ und „Filterung von nur nicht-synonymen kodierenden Varianten“ eine gefilterte ROI bestimmt.

2.1.3 Die einfache Region von Interesse

Die einfache ROI (bROI) ist im Wesentlichen dadurch definiert, dass sie eine abgeschlossene funktionelle Einheit bildet. Diese funktionelle Einheit kann beispielsweise aus einem Gen bestehen, das ein bestimmtes Protein kodiert, aber auch aus einer Gengruppe, die beispielsweise für einen komplexen Stoffwechselprozess verantwortlich ist. Gemäß der RVCD-Theorie kann das Vorhandensein bereits einer Variante innerhalb dieser funktionellen Einheit bei einer Person krankheitsverursachend sein. Daher ist für die Definition der bROI die Lokalisierung innerhalb des Genoms ausschlaggebend (Byrnes et al. 2013). Um eine funktionelle Einheit lokalisieren zu können, muss

insbesondere die Chromosomennummer und die Basenpaarposition für jede einzelne Variante innerhalb der ROI bekannt sein. Da sich das menschliche Genom aus 23 Paaren von homologen Chromosomen mit jeweils fortlaufenden Basen zusammensetzt, kann die Angabe der Chromosomennummer zusammen mit der Basenpaarposition als eine Art „Adresse“ der Variante innerhalb des Genoms angesehen werden. Diese Informationen werden in der Regel von dem Produzenten der Sequenzdaten in Form von Annotationsdateien mitgeliefert. Um schließlich die Zuweisung zu einer funktionellen Einheit durchführen zu können, werden im Allgemeinen Informationen aus externen Datenbanken wie *ANNOVAR* (Wang et al. 2010) benötigt. Diese Informationen werden mit Hilfe von Annotationsdatenbanken hinzugefügt, wie in Abbildung 2.2 dargestellt. Dabei ist zu beachten, dass die Definition einer bROI über eine funktionelle Einheit dazu führt, dass automatisch Varianten aus der Analyse ausgeschlossen werden, die keiner funktionellen Einheit zugewiesen werden können, wie z.B. intergenische Varianten, also Varianten, die sich zwischen zwei Genen befinden.

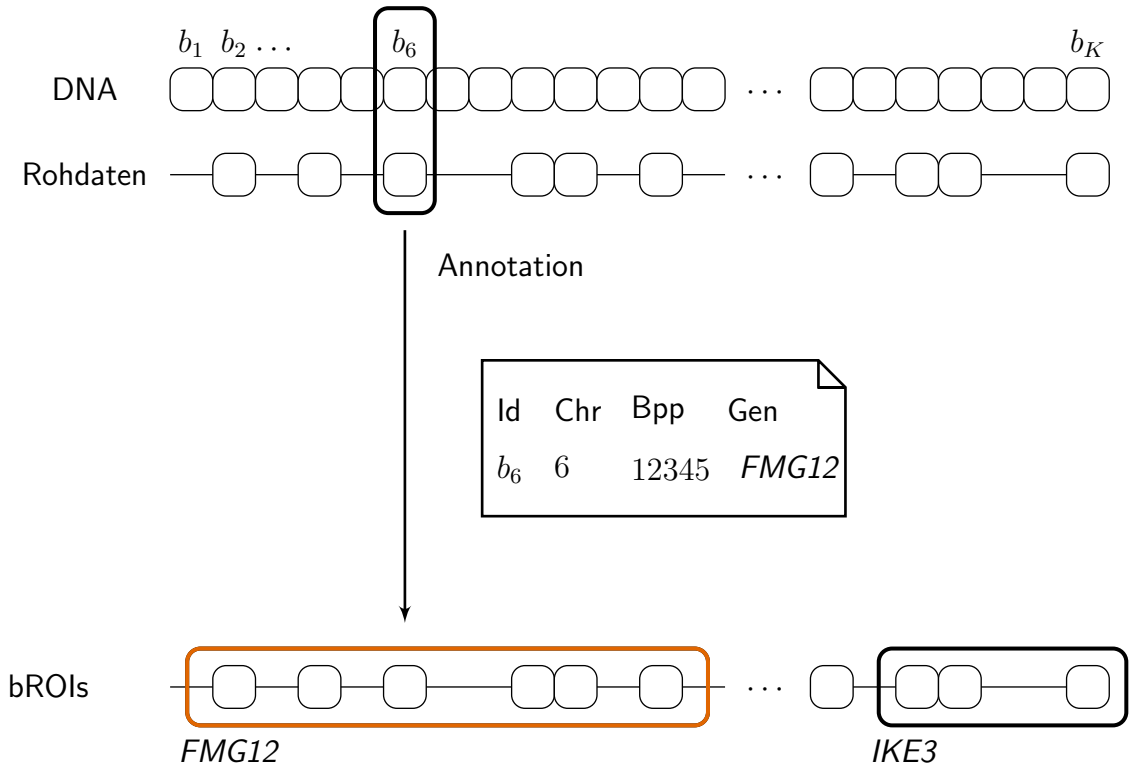


Abbildung 2.2: Bestimmung einer einfachen Region von Interesse (engl. basic region of interest(bROI)) aus Sequenzdaten einer Stichprobe durch hinzufügen von Chromosomennummer und Basenpaarposition sowie durch Zuweisung zu einer funktionellen Einheit.

2.1.4 Die gefilterte Region von Interesse

Ausgehend von der einfachen ROI kann eine noch weiter spezifizierte ROI, die gefilterte Region von Interesse (fROI), gebildet werden. Dabei werden zusätzliche kategorielle oder stetige Filterkriterien zur einfachen Region von Interesse hinzugefügt, nach denen dann gefiltert wird (Abbildung 2.3). Die zusätzliche Filterung kann aus

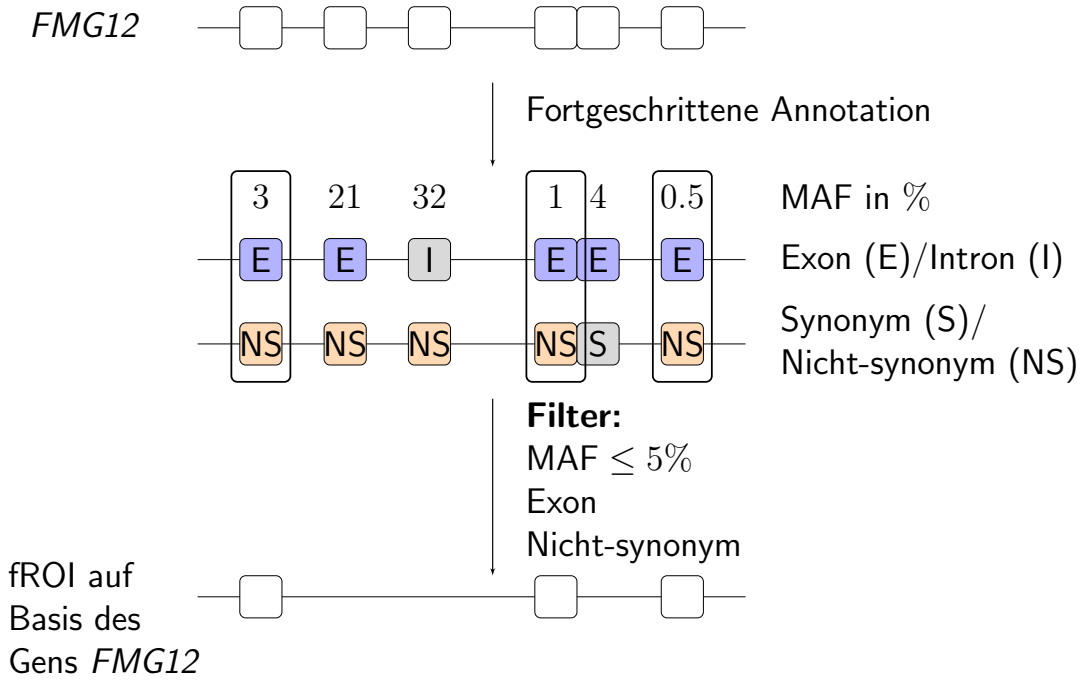


Abbildung 2.3: Schematische Darstellung zur Bestimmung der gefilterten Region von Interesse für eine gegebene Stichprobe und das fiktive Gen *FMG12*. Filterkriterien sind: Minor Allele Frequency (MAF) $\leq 5\%$, nicht synonyme (NS) Varianten aus Exon-Abschnitten.

unterschiedlichen Gründen sinnvoll sein. Möglicherweise liegt aus vorangegangenen Studien Vorwissen vor, so dass die Untersuchung auf eine Assoziation mit einem Phänotyp auf bestimmte Regionen eingegrenzt werden kann. Des Weiteren ist meist unbekannt, ob die Varianten in einer ROI verschiedene Effektstärken oder -richtungen haben. Da viele klassische Gruppierungsmethoden einen gemeinsamen Effekt der Varianten einer ROI annehmen (Derkach et al. 2014; Lee, Emond et al. 2012), kann es passieren, dass neutrale oder bi-direktionale Varianten das Assoziationssignal stören, was wiederum zu einem Verlust der Teststärke führen kann (Lee et al. 2014). Daher ist es insbesondere sinnvoll Informationen über die Art der Funktionalität einer Variante zu sammeln, so dass danach gefiltert werden kann. Typische Filterkriterien wurden in Abschnitt 2.1.2 beschrieben.

2.2 Eigenschaften von Gruppierungsmethoden

In diesem Abschnitt werden Probleme betrachtet, die bei der Untersuchung oder der Gruppierung seltener Varianten auftreten können. Aus den Lösungsansätzen für diese Probleme haben sich häufig die Grundidee bzw. die definierenden Eigenschaften für eine neue Gruppierungsmethode ergeben. Die Eigenschaften der Gruppierungsmethoden werden im Folgenden im Hinblick auf zwei Aspekte untersucht: Zum einen werden ihre Eigenschaften aus technischer Sicht, d.h. aus Sicht der zugehörigen Teststatistiken präsentiert und diskutiert. Dazu gehören die Interpretation der Genotypdaten innerhalb der Teststatistik, die Schätzung des p-Werts, die Arten der zu untersuchenden Phänotypen sowie die Einbindung möglicher Gewichte und Kovariablen. Zum anderen werden die Probleme bzw. Eigenschaften bzgl. ihrer biologischen Relevanz beleuchtet und eingeordnet. Hierzu zählen die mutmaßliche Funktion der Varianten, die Relevanz nicht genetischer Kovariablen und die parallele Untersuchung seltener und häufiger Varianten. Zu diesem Zweck wird jede Eigenschaft bzw. jedes Problem einzeln definiert und wenn möglich in verschiedenen Ausprägungen darstellt. Zudem erfolgt jeweils eine Interpretation im Kontext der RVCD-Theorie.

2.2.1 Kodierung der Region von Interesse

Nachdem eine ROI definiert wurde (Abschnitt 2.1), kann diese mit Hilfe einer Gruppierungsmethode auf eine mögliche Assoziation mit einem Phänotypen hin untersucht werden. Auf Basis der RVCD-Theorie haben sich für die Weiterverarbeitung der Genotypen der Varianten einer ROI verschiedene Möglichkeiten entwickelt. Diese lassen sich im Wesentlichen zwei Gruppen, *Indikatorkodierungen* und *Genotypkodierungen* zuordnen. Jede Kodierung basiert auf der folgenden Darstellung der Genotypen der Varianten $j = 1, \ldots, K$ der ROI für die Individuen $i = 1, \ldots, n$ der betrachteten Stichprobe:

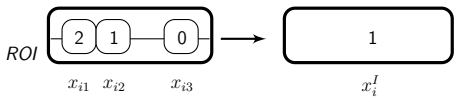
$$x_{ij} = \begin{cases} 0, & aa, \\ 1, & Aa, \\ 2, & AA, \end{cases} \qquad i = 1, \ldots, n, \quad j = 1, \ldots, K. \tag{2.1}$$

Dabei ist a das häufige und A das seltene Allel. Trägt z.B. das Individuum i zwei Kopien des seltenen Allels der j-ten Variante, wird der Genotyp mit $x_{ij} = 2$ kodiert.

2.2.1.1 Indikatorkodierung

Die frühen Gruppierungsmethoden basieren fast ausschließlich auf der Indikatorkodierung. Diese greift die ursprüngliche Annahme der RVCD-Theorie auf, dass alle Varianten einer ROI die gleiche Effektrichtung wie auch eine gemeinsame Effektstärke haben. Dabei wird für eine vorliegende ROI für jedes Individuum i lediglich geprüft, ob mindestens ein seltenes Allel in der Region vorliegt oder nicht. Dementsprechend wird die Gruppe der Varianten bzw. Genotypen der ROI in eine Indikatorvariable x_i^{I} umkodiert:

$$x_i^{\mathrm{I}} := \begin{cases} 1, & \sum_{j=1}^{K} x_{ij} > 0, \\ 0, & \text{sonst.} \end{cases}$$

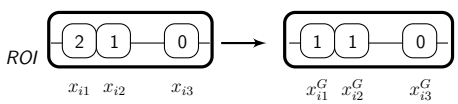


Diese hat für das Individuum i den Wert 1, wenn mindestens ein seltenes Allel in der ROI enthalten ist, ansonsten ist sie 0. Die Stärke der Indikatoransatzes besteht darin, dass die RVCD-Theorie direkt auf die Kodierung der ROI übertragen wird.

2.2.1.2 Genotypkodierung

Die neueren Gruppierungsmethoden gehen von individuellen Effektstärken und -richtungen der Varianten innerhalb einer ROI aus. Bei der Genotypkodierung wird zwar ebenso die Gruppe der Varianten insgesamt betrachtet, allerdings fließt jede Variante einzeln in die entsprechende Teststatistik ein. Da bei der RVCD-Theorie ausschließlich seltene Varianten berücksichtigt werden, ist insbesondere das Vorkommen einer zweifachen Kopie eines seltenen Allels (der Fall AA) selten, so dass sich die Kodierung aus Gleichung (2.1) in den meisten Fällen reduziert:

$$x_{ij}^{\mathrm{G}} := \begin{cases} 1, & x_{ij} > 0, \\ 0, & \text{sonst.} \end{cases} \qquad (2.2)$$



2.2.2 Putative Funktion von Varianten

Eine genetische Variante kann, egal ob häufig oder selten, vereinfacht ausgedrückt drei mögliche Wirkungen auf die Ausbildung eines betrachteten körperlichen Merkmals bzw. Phänotyps haben: neutral, schädigend oder protektiv. Sind zum Beispiel höhere Werte eines quantitativen Merkmals gleichbedeutend mit einem verschlechterten Gesundheitszustand, dann hat die Präsenz einer schädigenden Variante höhere Werte des Merkmals zur Folge. Eine protektive Variante kann hingegen geringere Werte dieses Merkmals hervorrufen. Im Allgemeinen treten schädigende Wirkungen häufiger auf als schützende (Han und Pan 2010). Varianten ohne Wirkung auf den Phänotyp treten am häufigsten auf (Kryukov et al. 2007), die beiden übrigen Möglichkeiten sind wesentlich seltener. Zu beachten ist aber, dass dies insbesondere für häufige Varianten gilt. Denn die Untersuchungen von Zhu et al. (2011) deuten darau fhin, dass die funktionelle Relevanz bzw. die (negative) Auswirkung einer Variante umgekehrt proportional zur MAF steigt, sobald ihre MAF unter $8\% - 10\%$ sinkt.

In verschiedenen Beispielen konnten für seltene Varianten eines Gens auch bi-direktionale Effekte nachgewiesen werden (Cohen et al. 2004; Fitze et al. 2002). Es konnte gezeigt werden, dass sich seltene Varianten bestimmter Gene entweder in einem oder in beiden Extremen der Verteilung des Phänotyps der Stichprobe häufen. So ist das Gen *PCSK9* ein Beispiel für eine Gruppe von seltenen Varianten, die sowohl mit hohen als auch niedrigen Werten des Lipoproteins niederer Dichte (engl. Low Density Lipoprotein, LDL) im Zusammenhang steht (Kotowski et al. 2006; Zhang et al. 2011).

Eine bestehende Assoziation einer heterogen wirkenden Gruppe von Varianten mit einem Phänotyp kann sehr schwer zu identifizieren sein. Dies gilt insbesondere bei Methoden, die auf der Indikatorkodierung basieren (Abschnitt 2.2.1.1), bei der angenommen wird, dass die Varianten einer ROI alle kausal sind und hinsichtlich der Ausprägung des Phänotyps mit derselben Effektstärke in eine Richtung wirken (Lee, Emond et al. 2012). Dennoch ist es möglich, dass innerhalb einer ROI Varianten mit unterschiedlicher Effektstärke und -richtung auftreten (Kimura 1968; Wu et al. 2011a). Dann kann es bei Verwendung der Indikatorkodierung passieren, dass heterogene Effekte in der Region von Interesse völlig verwischt werden, so dass weder der Effekt selbst noch die bi-direktionale Wirkung der Varianten erkannt werden. Gibt es vor

der Assoziationsuntersuchung eines Merkmals und einer ROI keine Hinweise über die Art der Effektstruktur kann die Betrachtung der Individuen der Stichprobe hilfreich sein, die sich in den Extremen der Mermalsverteilung befinden. Ist dies nicht möglich weil zum Beispiel die Assoziation zum Fall-Kontroll-Status untersucht wird, kann es sinnvoll sein, geeignete Gruppierungsmethoden in Betracht zu ziehen. Diese sollten zum einen die Genotypkodierung verwenden und zum anderen eine mögliche bi-direktionale Wirkung von Varianten über variantenweise Gewichte oder paarweise Vergleiche von Individuen der Stichprobe berücksichtigen wie der *Sequencing Kernel Association Test* (*SKAT*) (vgl. Abschnitt 2.3.14). Aus der Unterscheidung der angenommenen Effektstruktur innerhalb der ROI haben sich zwei Klassen von Gruppierungsmethoden etabliert, die *burden* und *non-burden* Tests. Dabei wird der Begriff „burden“ (dt.: Last) für die Methoden verwendet, die von einem gemeinsamen Effekt aller Varianten einer ROI in die gleiche Richtung ausgehen (Lee, Emond et al. 2012; Neale et al. 2011; Wu et al. 2011a). Dementsprechend werden die Methoden, die auch heterogene Effekte innerhalb der ROI berücksichtigen als non-burden bezeichnet.

2.2.2.1 Fehlklassifizierung

Bei den Annahmen über die Funktion einer Variante innerhalb einer ROI sind zwei grundsätzliche „Fehler“ zu unterscheiden: Eine

Funktionelle Fehlklassifizierung (FF) tritt auf, wenn neutrale Varianten in der ROI enthalten sind oder wenn funktional relevante Varianten in der ROI fehlen.

Direktionale Fehlklassifizierung (DF) tritt auf, wenn angenommen wird, dass die Varianten einer ROI mit derselben Effektstärke in die gleiche Richtung wirken, obwohl bi-direktionale Effekte verschiedener Stärke innerhalb der Region von Interesse vorliegen.

2.2.2.2 Lösungsansätze bei Fehlklassifizierungen

Eine FF kann verschiedene Ursachen haben. Zum einen können Informationen über relevante Varianten auf Grund einer unzureichenden Sequenzierungsqualität fehlen,

zum anderen können vorhandene Varianten in der Annotation durch Vorhersage-Tools fälschlich als *neutral* eingestuft und somit gefiltert werden, obwohl sie nicht neutral sind. Dann kann eine zuverlässige Annotation der Varianten hinsichtlich ihrer funktionellen Relevanz durch Programme wie *PolyPhen2* oder *SIFT* hilfreich sein um die FF einzuschränken.

Ein weiterer Ansatz zur Vermeidung von FF stützt sich auf die Annahme, dass es eine optimale Frequenz des seltenen Allels gibt, so dass unterhalb dieser Frequenz alle Varianten einer Region von Interesse als funktionell relevant angenommen werden können (Manolio 2010). Diese Annahme liegt u.a. in der natürlichen Selektion begründet und diente als Basisidee für die Entwicklung der *Variable Threshold* (*VT*) Gruppierungsmethode von Price et al. (2010) (Abschnitt 2.3.8).

Eine DF ist wesentlich schwieriger zu umgehen, da hier keine weiteren Lösungsansätze existieren außer der Wahl einer geeigneten Gruppierungsmethode, die robust gegenüber DF ist (also ein non-burden Test). Ob und in welche Richtung eine Variante wirkt, lässt sich ohne eine vorangegangene (Regressions-) Analyse oft nicht sagen. Um mögliche bi-direktionale Effekte detektieren zu können, verwenden Methoden wie der *Sequencing Kernel Associaton Test* (*SKAT*) oder *optimal unified SKAT* (*SKAT-O*) (Abschnitte 2.3.14 und 2.3.15) häufig Daten-abhängige Gewichte (Abschnitt 2.2.4.4) und betrachten den Einfluss der einzelnen Varianten in der ROI, indem sie je Variante die Individuen paarweise miteinander vergleichen.

2.2.3 Gemeinsame Berücksichtigung seltener und häufiger Varianten

Die Frage ob eine oder mehrere häufige Varianten mit moderaten Effekten (CVCD) oder eine Gruppe seltener Varianten mit starken Penetranzen (RVCD) krankheitsverursachend ist, wurde in der Literatur intensiv diskutiert (Li und Leal 2008; Schork et al. 2009). Die Penetranz ist definiert als der Anteil von Individuen einer Stichprobe, die Träger einer oder mehrerer bestimmter genetischer Varianten sind und einen betrachteten Phänotyp ausgeprägt haben (Katsanis und Katsanis 2013). Beide Theorien konnten bereits in zahlreichen Studien belegt werden. Jedoch waren beide (allein) bei vielen Krankheiten nicht in der Lage, die vollständige Ätiologie, also die

Ursachen und die Entstehung einer Krankheit, zu erklären (Cohen et al. 2004; Fitze et al. 2002; Kotowski et al. 2006; Manolio 2010; Seng und Seng 2008; Speliotes et al. 2011).

Aus diesem Grund entwickelte sich eine neue Diskussion, die sich insbesondere mit dem MAF-Spektrum der krankheitsverursachenden Varianten einer zu untersuchenden Krankheit beschäftigte (Manolio et al. 2009; Pritchard 2001; Schork et al. 2009). Schließlich entwickelte sich eine weitere Theorie, die Interaktionen von häufigen und seltenen Varianten für die Krankheitsbildung in Betracht zieht (Bodmer und Bonilla 2008). In diesem Zusammenhang sind insbesondere zwei Arten von Interaktionen von Bedeutung. Zum einen ein verstärkender Effekt von seltenen Varianten und zum anderen ein Einfluss von häufigen auf seltene Varianten. Ist eine häufige Variante hinsichtlich der Assoziation mit einem Phänotyp bereits identifiziert worden, so kann ihr Effekt durch seltene Varianten aus der gleichen Region verstärkt werden, was extreme Ausprägungen des Phänotyps verursachen kann. Eine Bestätigung dieser Theorie wurde für den Methotrexat-Abbaus bzgl. des Gens *SLCO1B1* durch Ramsey et al. (2012) gezeigt. Ein anderes Beispiel stellt ist das Liddle Syndrom, eine seltene Form des Bluthochdrucks die durch seltene Varianten beeinflusst wird (Schork et al. 2009). Genauso könnten häufige Varianten einen Einfluss auf den Effekt von seltenen Varianten haben (Bodmer und Bonilla 2008; Felix et al. 2006). Diese Interaktions-Theorie konnte in einigen Beispielen untermauert werden. Beispielsweise konnte gezeigt werden, dass häufige Varianten in den Genen *GSTT1* und *GSTM1* bei Patienten mit dem Lynch-Syndrom zusammen mit einer seltenen nonsense-Variante die Lage, den Zeitpunkt des Auftretens und andere Symptome der Krebserkrankung beeinflussen. Ähnlich wird offenbar die erbliche Form des Brustkrebses zusätzlich durch seltene, hoch-penetrante Varianten der Gene *BRCA1* und *BRCA2* beeinflusst (Bodmer und Bonilla 2008; Schork et al. 2009).

Die Mehrzahl der hier untersuchten Gruppierungsmethoden basiert ausschließlich auf der Untersuchung von seltenen Varianten. Nur wenige Gruppierungsmethoden wie der *Combined and Multivariate Collapsing* (*CMC*) Test von Li und Leal (2008) oder der *Sequencing Kernel Association Test* (*SKAT*) von Wu et al. (2011a) lassen von vornherein eine gemeinsame Untersuchung seltener und häufiger Varianten zu.

2.2.4 Regressionsansatz

Die Regression ist ein Standardverfahren der Statistik (Mendenhall et al. 1996) und wird u.a. im Kontext von GWAS genutzt, um eine Vorhersage über eine abhängige Variable (Phänotyp) hinsichtlich der Ausprägung einer unabhängigen (erklärenden) Variable (z.B. den Genotyp einer Variante) treffen zu können. Die abhängige Variable kann von weiteren, möglicherweise unbekannten oder nicht erfassbaren Variablen wie Umweltbedingungen oder Ausprägungen weiterer Phänotypen beeinflusst werden. Diese Schwankungen werden durch einen stochastischen Fehlerterm in die Vorhersagefunktion einbezogen. Ändert sich die abhängige Variable proportional zur Änderung der unabhängigen Variablen handelt es sich um eine lineare Regression und die Regressionsfunktion kann durch eine Gerade dargestellt werden. Mittels Regression lässt sich bestimmen, wie groß der Effekt einer unabhängigen Variable auf eine abhängige Variable ist. Basiert das Modell auf nur einer unabhängigen Variablen, so spricht man von einer einfachen linearen Regression. Werden hingegen mehrere unabhängige Variablen im Modell betrachtet, um eine abhängige Variable vorherzusagen, handelt es sich um sogenannte multiple oder Mehrfach-Regression. In diesem Fall können weitere Informationen in Form von Kovariablen für die Vorhersage der abhängigen Variablen mit in das Modell einbezogen werden.

Die meisten der in dieser Arbeit betrachteten Gruppierungsmethoden lassen sich durch ein Regressionsmodell beschreiben. Daher soll in diesem Abschnitt ein allgemeines Regressionsmodell unter Einbeziehung von Kovariablen aufgestellt werden, das leicht an jede der vorgestellten Gruppierungsmethoden angepasst werden kann. Die Untersuchung stetiger (quantitativer) und binärer Merkmale bzw. Phänotypen wird mit Blick auf die Besonderheiten bei der Betrachtung seltener Varianten diskutiert. Des Weiteren wird auf die Berücksichtigung von Kovariablen bei der Untersuchung seltener Varianten eingegangen. Arten von Gewichten im Kontext von Gruppierungsmethoden und ihr Einfluss auf die Leistung der Ansätze hinsichtlich der Teststärke und des Fehlers erster Art werden im Anschluss diskutiert.

2.2.4.1 Allgemeines Regressionsmodell

Es seien die Genotypen $\boldsymbol{X} = (x_{ij})_{\substack{i=1,\dots,n,\\ j=1,\dots,K}}$ einer Stichprobe von n Individuen mit je K Varianten einer ROI, die Ausprägung eines normalverteilten Phänotyps $\boldsymbol{y} = (y_1, \dots, y_n)^\intercal$ und Kovariablen $\boldsymbol{C} = (c_{iq})_{\substack{i=1,\dots,n,\\ q=1,\dots,Q}}$ gegeben. Dann lässt sich das allgemeine Regressionsmodell für das Individuum i wie folgt definieren:

$$\begin{aligned} y_i &= \beta_0 + (\beta_1, \dots, \beta_K) \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & w_K \end{pmatrix} \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iK} \end{pmatrix} + (\alpha_1, \dots, \alpha_Q) \begin{pmatrix} c_{i1} \\ \vdots \\ c_{iQ} \end{pmatrix} + \varepsilon_i \\ &= \beta_0 + \boldsymbol{\beta}^\intercal \boldsymbol{W} \boldsymbol{x}_i + \boldsymbol{\alpha}^\intercal \boldsymbol{c}_i + \varepsilon_i. \end{aligned} \tag{2.3}$$

Dabei ist β_0 der sog. Achsenabschnitt, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^\intercal$ der Effektvektor, $\boldsymbol{w} = (w_1, \dots, w_K)^\intercal$ die Hauptdiagonale der Diagonalmatrix $\boldsymbol{W}$ und $\boldsymbol{x}_i = (x_{i1}, \dots, x_{iK})^\intercal$ ist der Genotyp-Vektor der K Varianten für das Individuum i. Die Kovariablen und die zugehörigen Regressoren sind durch $\boldsymbol{c_i} = (c_{i1}, \dots, c_{iQ})^\intercal$ und $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_Q)^\intercal$ gegeben. Der residuale Effekt für Person i wird mit ε_i bezeichnet und wird als normalverteilt mit Mittelwert $\mu = 0$ und Varianz σ^2 angenommen.

Zu jeder Regressions-basierten Gruppierungsmethode existiert ein analoges Modell, in dem ein dichotomer (binärer) Phänotyp betrachtet werden kann. In dem zugehörigen Modell wird der Phänotyp mittels einer logit-Funktion wie in Gleichung (2.4) angepasst, während die unabhängigen Variablen analog zu Gleichung (2.3) definiert sind:

$$\begin{aligned} \operatorname{logit} P(y_i = 1) &= \ln\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) \\ &= \beta_0 + \boldsymbol{\beta}^\intercal \boldsymbol{W} \boldsymbol{x}_i + \boldsymbol{\alpha}^\intercal \boldsymbol{c}_i. \end{aligned} \tag{2.4}$$

Die logit-Funktion wird dazu verwendet, um Verhältnisse von Wahrscheinlichkeiten und den entsprechenden Gegenwahrscheinlichkeiten für ein bestimmtes (diskretes) Ereignis wie dem Krankheitsstatus zu linearisieren. Dabei ist ln der natürliche Logarithmus zur Basis e (Eulersche Zahl). Weiterhin ist β_0 der sog. Achsenabschnitt, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^\intercal$ der Effektvektor, $\boldsymbol{w} = (w_1, \dots, w_K)^\intercal$ die Hauptdiagonale der

Diagonalmatrix $\boldsymbol{W}$ und $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK})^\intercal$ ist der Genotyp-Vektor der K Varianten für das Individuum i. Die Kovariablen und die zugehörigen Regressoren sind durch $\boldsymbol{c_i} = (c_{i1}, \ldots, c_{iQ})^\intercal$ und $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)^\intercal$ gegeben.

Da zu einer ROI in der Regel mehrere Varianten gehören, wird im Allgemeinen zunächst ein multiples Regressionsmodell angesetzt. Dieses Modell kann sich bei Gruppierungsmethoden, die die Information der Gruppe von Varianten mittels Indikatorkodierung zu einer sog. Super-Variante kumulieren, auf ein einfaches Regressionsmodell mit einem kollektiven Effekt (Abschnitt 2.2.4.4) reduzieren. Das Regressionsmodell reduziert sich außerdem, unabhängig von Kodierung, wenn keine Kovariablen in die Vorhersage des Phänotyps miteinbezogen werden oder nicht verfügbar sind.

2.2.4.2 Art des Phänotyps

Ein Phänotyp kann verschieden skaliert sein und man unterscheidet dabei qualitative und quantitative Phänotypen. Während sich qualitative Phänotypen in diskrete Kategorien einteilen lassen, die einander ausschließen, sind quantitative Phänotypen hingegen stetige, oft messbare Merkmale. Ein Spezialfall stellt der dichotome (binäre) Phänotyp dar, der genau zwei Ausprägungen aufweist. Ein Beispiel hierfür ist der Fall-Kontroll-Status, der Teilnehmer einer Studie in „krank“ und „gesund“ bzw. „1“ und „0“ einteilt.

Die meisten Gruppierungsmethoden sind so konzipiert, dass auf die Assoziation einer ROI mit einem dichotomen Phänotyps wie dem Fall-Kontroll-Status hin untersucht wird. Es gibt aber auch einige Ansätze, insbesondere die Regressions-basierten, die zur Assoziationsuntersuchung neben binären auch quantitative Phänotypen zulassen. In Regressionsanalysen ist im Allgemeinen die Betrachtung von quantitativen der von binären Phänotypen vorzuziehen, da so das gesamte Variationsspektrum des Phänotyps in die Modellbildung eingeht, was zu besseren Ergebnissen bzgl. der Teststärke führt.

Bei der Analyse von seltenen Varianten hinsichtlich quantitativer Phänotypen kann angenommen werden, dass die krankheitsverursachenden Varianten, insbesondere die

Funktionsverlustvarianten (engl. Loss-Of-Function Variant), vor allem bei den Individuen zu finden sind, die eine extreme Ausprägung des Phänotyps aufweisen (Bacanu et al. 2011; Manolio et al. 2009). Diese These konnte in zahlreichen Arbeiten bestätigt werden (Cohen et al. 2004; Kotowski et al. 2006; Ramsey et al. 2012; Risch und Zhang 1995). Barnett et al. (2013) haben hierfür aus einer Stichprobe eine Teilstichprobe aus den Extremen der Phänotypverteilung (engl. Extreme phenotype sampling, EPS) untersucht und konnten nachweisen, dass sich krankheitsverursachende seltenen Varianten insbesondere in den Extremen der Phänotypverteilung befinden. Peloso et al. (2015) konnten außerdem zeigen, dass bei Verwendung von EPS für seltene Varianten bessere Ergebnisse in der Teststärke erzielt werden als bei häufigen Varianten. Bei der Verwendung von Gruppierungsmethoden, die lediglich eine Untersuchung von binären Phänotypen zulassen, kann eine Dichotomisierung eines stetigen Phänotyps, also die Zuordnung stetiger Merkmalsausprägungen zu zwei einander ausschließenden Gruppen, durchaus eine gültige Methode sein, wenngleich die Teststärke möglichweise geringer ist als bei Ansätzen, die die Betrachtung des ursprünglichen quantitativen Phänotyps zulassen (Barnett et al. 2013).

2.2.4.3 Kovariablen

Meist ist es bei der Assoziationsanalyse eines Phänotyps wie dem Krankheitsstatus mit einer oder mehrerer genetischer Mutationen stets sinnvoll, auch Kovariablen wie Alter, Geschlecht oder die Populationszugehörigkeit in die Analyse mit einzubeziehen (Stitziel et al. 2011). Insbesondere bei seltenen Varianten handelt es sich oft um Varianten, die nur eine bis wenige Generationen weitervererbt wurden oder sogar neu sind (*de novo Mutation*). Daher ist hier besonders die simultane Untersuchung der ROI mit zusätzlichen Kovariablen wie der Populationszugehörigkeit von großer Bedeutung (Lin und Tang 2011). Bansal et al. (2010) empfehlen bei Assoziationsanalysen im Zusammenhang mit seltenen Varianten die Verwendung eines umfassenden Regressionsansatzes, um auf mögliche Kovariablen oder Verzerrungseffekte adjustieren zu können. Insbesondere die Verwendung von häufigen Varianten als Kovariable, für die bereits eine Assoziation zum Phänotyp des Interesses nachgewiesen werden konnte, stellen die Autoren als wichtig heraus, um die Teststärke der Gruppierungsmethoden zu erhöhen.

Die meisten frühen Gruppierungsmethoden wurden nicht für eine Berücksichtigung von Kovariablen konzipiert. Unter ihnen ermöglichen lediglich die Ansätze von Morris und Zeggini, die ersten Regressionsansätze unter den Gruppierungsmethoden, die Berücksichtigung von Kovariablen (Morris und Zeggini 2010). Allerdings ist es prinzipiell leicht möglich, die meisten Regressions-basierten Ansätze entsprechend zu erweitern. Die jüngeren Gruppierungsansätze folgen der Empfehlung von Bansal et al. (2010) und sind komplexe Regressionsansätze (Bansal et al. 2010; Lee, Emond et al. 2012; Wu et al. 2011a).

2.2.4.4 Gewichte

Aufgrund der geringen MAF ist die Teststärke gewöhnlicher Assoziationstests bei der Anwendung auf seltene Varianten sehr gering. Daher nutzen viele Gruppierungsmethoden spezielle zusätzliche Gewichte um den Effekt der Varianten innerhalb einer ROI zu stärken oder abzuschwächen. Byrnes et al. (2013) geben an, dass bei der Analyse seltener Varianten der Einsatz von Gewichten das entscheidende Mittel ist, um überhaupt vorhandene Assoziationen zu finden.

Im Kontext von Regressions-basierten Gruppierungsmethoden kann der Begriff des Gewichts irreführend sein. Denn sowohl der Effektschätzer (im Allgemeinen $\boldsymbol{\beta}$) als auch ein zusätzlicher individueller Parameter (w_j für eine Variante $j \in \{1, \ldots, K\}$) können ein Gewicht für eine bestimmte Variante in der ROI darstellen. Auch eine Kombination aus beidem ist, wie in den allgemeinen Modellen der Gleichungen (2.3) und (2.4) dargestellt möglich. Daraus ergeben sich im Wesentlichen zwei Möglichkeiten eine ROI bzw. die darin enthaltenen Varianten zu gewichten. Zum einen können alle Varianten einer ROI dasselbe Gewicht erhalten, was insbesondere dann sinnvoll ist, wenn die Annahme eines kollektiven Effekts der Varianten in ein und dieselbe Richtung besteht, wie in den Ansätzen von Morgenthaler und Thilly (2007) oder Morris und Zeggini (2010). Zum anderen kann jede Variante ein individuelles Gewicht erhalten und die Genotypinformation der Varianten mit den entsprechenden Gewichten akkumuliert und zusammen analysiert werden (Hoffmann et al. 2010; Madsen und Browning 2009; Pongpanich et al. 2011; Zhang et al. 2011).

Beide Ansätze zur Gewichtung von seltenen Varianten gehen auf biologische Theorien zurück. Eine Theorie besagt, dass der Effekt der Variante anti-proportional zu dessen zugehöriger MAF ist (Gorlov et al. 2008; Manolio et al. 2009). Daher schlagen Basu und Pan (2011) und Madsen und Browning (2009) in ihrer Methode ein individuelles Gewicht für jede Variante vor, was den Genotyp der Variante mit der Varianz der entsprechenden MAF in der Gruppe der Kontrollen normiert. Auf diese Weise wird der Effekt einer schädlichen Variante im Gegensatz zu einer neutralen Variante hervorgehoben (Basu und Pan 2011; Madsen und Browning 2009). Derkach et al. (2014) stellen diesen Ansatz jedoch in Frage, da es im Falle einer protektiven und einer neutralen Variante mit der gleichen MAF zu einer Herabsetzung des Gewichts der protektiven im Gegensatz zur neutralen Variante kommt. Hoffmann et al. (2010) bestätigen, dass Gruppierungsansätze, deren Gewichte auf der MAF basieren im Falle von zweiseitigen Effekten innerhalb einer Region von Interesse nicht zuverlässig sind.

Des Weiteren gibt es Gruppierungsmethoden, die eine Kombination aus individuellen Varianten-Effekten und einem gemeinsamen Effekt der gesamten ROI in der Assoziationsanalyse verwenden. So schätzen Han und Pan (2010) Daten-adaptiv in einem zweistufigen Regressionsansatz zunächst die individuelle Effektrichtung und lassen diese Information in ein zweites Regressionsmodell einfließen, in dem der gemeinsame Effekt der gesamten ROI geschätzt wird. Auch Zhang et al. (2011) nutzen in ihrem Ansatz ebenfalls ein zweistufiges Verfahren, in dem zunächst die individuellen Effekte der Varianten innerhalb einer ROI geschätzt werden. Aus den entsprechenden individuellen linksseitigen p-Werten werden dann Varianten-spezifische Gewichte gebildet. Diese dienen als individuelle Gewichte der Genotypen der Varianten in einem weiteren Regressionsmodell, welches den kollektiven Effekt der ROI schätzt.

Einige Autoren schlagen auch die Verwendung von Wahrscheinlichkeiten für die vorhergesagte Schädlichkeit einer Variante mittels *SIFT*, *PolyPhen2* o.ä. vor (Asimit und Zeggini 2009; Price et al. 2010; Zhang et al. 2011). Weitere Möglichkeiten individuelle Gewichte zu schätzen sind eine Hauptkomponentenanalyse (Pongpanich et al. 2011) oder eine Fourier-Transformation, wie sie Luo et al. (2012) in ihrem Ansatz nutzen.

Obwohl durch die Verwendung von Gewichten in Gruppierungsmethoden die Teststärke wesentlich erhöht werden kann, ist ihr Einfluss begrenzt. Das liegt daran, dass selbst sehr kleine individuelle Gewichte für nicht-verursachende Varianten den detektierbaren Effekt von verursachenden Varianten innerhalb einer ROI verringern können (Bhatia et al. 2010).

2.2.5 Struktur der Teststatistik

Gruppierungsmethoden zur Analyse von seltenen Varianten wurden von Derkach et al. (2014) in zwei Klassen unterteilt, die hinsichtlich ihrer Teststärke miteinander verglichen wurden. Diese Klassen lassen sich an der verwendeten Teststatistik erkennen, die entweder ausschließlich lineare oder quadratische Effekte der Varianten betrachten. Ein Effekt ist linear wenn sich die Ausprägung des betrachteten Phänotyps proportional zur Präsenz einer oder mehrerer Varianten in der ROI verhält. Ein Effekt ist quadratisch wenn Varianten innerhalb einer Region von Interesse paarweise interagieren und somit eine komplexe Effektstruktur vorliegt. Diese kann durch geeignete Teststatistiken abgebildet werden. Im Folgenden werden deswegen sowohl der Unterschied in der Struktur der Teststatistik anhand eines quantitativen Phänotyps ohne Betrachtung von Kovariablen als auch die Leistungsunterschiede bei bestimmten Untersuchungsszenarien aus der Arbeit von Derkach et al. (2014) dargelegt.

2.2.5.1 Lineare Teststatistiken

Allen Gruppierungsmethoden mit linearen Teststatistiken ist gemein, dass variantenspezifische Scores S_j über alle Individuen einer Stichprobe gebildet werden. Anschließend werden diese Scores S_j aller Varianten der ROI für die Bildung der linearen Teststatistik T_L aufsummiert:

$$S_j = \sum_{i=1}^{n} (y_i - \hat{y}) x_{ij}, \quad j = 1, \ldots, K, \quad \text{mit } \hat{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{2.5}$$

$$T_L = \sum_{j=1}^{K} w_j S_j \tag{2.6}$$

Dabei sind w_j, $j = 1, \ldots, K$ die Gewichte der einzelnen Varianten wie im Regressionsmodell in Gleichung (2.3), y_i ist die Phänotypausprägung für das Individuum $i \in \{1, \ldots, n\}$ und x_{ij} ist der Genotyp von Individuum i für die Variante j.

Des Weiteren weisen alle Teststatistiken der linearen Klasse unter der Nullhypothese asymptotisch eine ähnliche Verteilung auf, nämlich eine χ^2-Verteilung mit einem Freiheitsgrad:

$$T_L^2 \sim \chi_1^2$$

Unter der Alternativhypothese H_1 lässt sich die asymptotische Verteilung der linearen Teststatistik als nicht-zentrale χ^2-Verteilung mit einem Freiheitsgrad und einem Nichtzentralitätsparameter ζ herleiten vgl. (Derkach et al. 2014):

$$T_L^2 \sim \chi_{1,\zeta}^2$$

Dabei hängt ζ von den verwendeten Gewichten in der Teststatistik, dem Erwartungswert und der Kovarianzmatrix des betrachteten Merkmals unter H_1 ab. Für eine detaillierte Angabe wird auf Derkach et al. (2014) verwiesen.

Beispiele für lineare Teststatistiken sind die Ansätze CAST von Morgenthaler und Thilly (2007), RVT1/2 von Morris und Zeggini (2010) und WSS von Madsen und Browning (2009).

2.2.5.2 Quadratische Teststatistiken

Quadratische Teststatistiken werden oft auch als Varianzkomponenten-Tests bezeichnet, und basieren auf Regressionsmodellen, haben aber eine komplexere Form als lineare Teststatistiken:

$$T_Q = \boldsymbol{S}^\intercal \boldsymbol{A} \boldsymbol{S} \tag{2.7}$$

Dabei ist $\boldsymbol{A}$ eine positiv definite oder semi-definite symmetrische Matrix, z.B. die Inverse der Kovarianzmatrix des Score-Vektors $\boldsymbol{S} = (S_1, \ldots, S_K)$ der Varianten der betrachteten ROI aus Gleichung (2.5). In Gleichung (2.7) kann man sehen, dass bei

den quadratischen Teststatistiken durch die Matrix $\boldsymbol{A}$ auch Interaktionseffekte der Varianten einer ROI erfasst werden.

Aufgrund der komplexeren Form quadratischer Teststatistiken sind auch die entsprechenden asymptotischen Verteilungen unter der Nullhypothese komplizierter. Sie ergeben sich als Linearkombination mehrerer χ^2-Verteilungen mit einem Freiheitsgrad:

$$T_Q \sim \sum_{j=1}^{P} \lambda_j \chi_1^2 \tag{2.8}$$

mit Koeffizienten $\lambda_1, \ldots, \lambda_P$, $P \leq K$. Unter der Alternativhypothese H_1 lässt sich die asymptotische Verteilung der quadratischen Teststatistik als Linearkombination nichtzentraler χ^2-Verteilungen mit einem Freiheitsgrad und Nichtzentralitätsparametern $\zeta_1, \ldots, \zeta_P$ herleiten:

$$T_Q \sim \sum_{j=1}^{P} \lambda_j \chi_{1,\zeta_j}^2. \tag{2.9}$$

mit Koeffizienten $\lambda_1, \ldots, \lambda_P$, $P \leq K$. Die Nichtzentralitätsparameter $\zeta_1, \ldots, \zeta_P$ hängen von Eigenvektoren der Matrix $\boldsymbol{A}$ und dem Erwartungswert des betrachteten Merkmals unter H_1 ab. Für eine detaillierte Angabe wird auf Derkach et al. (2014) verwiesen.

Die Ansätze SKAT von Wu et al. (2011a), SKAT-O von Lee, Emond et al. (2012) und C-α von Neale et al. (2011) sind Beispiele für quadratische Teststatistiken.

2.2.5.3 Vergleich linearer und quadratischer Teststatistiken

Derkach et al. (2014) zeigen, dass beide Klassen bei zu geringen Fallzahlen versagen. Mit steigender Fallzahl dominiert die quadratische die lineare Klasse. Erst bei einem Anteil von mehr als 75% verursachenden Varianten innerhalb der ROI erreicht die Teststärke beider Klassen Werte von über 80%. Die linearen Statistiken sind den quadratischen in kleinen Studien dann überlegen, wenn fast alle Varianten einen Effekt aufweisen, der in die gleiche Richtung wirkt. Ist ein Großteil der Varianten innerhalb der ROI neutral, können hingegen die quadratischen die linearen Test-

statistiken dominieren. Die Autoren kommen zu dem Schluss, dass die Teststärke aller Gruppierungsmethoden im Wesentlichen von drei Faktoren abhängt: dem Anteil der verursachenden Varianten, der Richtung der Assoziation der Varianten in der ROI zum untersuchten Phänotyp (schädlich, protektiv oder beides) und dem Verhältnis der Frequenz der seltenen Allele der betrachteten Varianten zu den entsprechenden genetischen Effekten. Diese Aussagen werden in Kapitel 3 anhand eines Simulationsdatensatzes überprüft.

2.2.6 Schätzung des p-Werts

Etwa für die Hälfte der hier betrachteten Gruppierungsmethoden ist die Verteilung der Teststatistik unter der Nullhypothese, dass keine Assoziation zwischen der betrachteten ROI und dem Phänotypen vorliegt, unbekannt. In diesem Fall wird die Verteilung unter der Nullhypothese und somit der gesuchte p-Wert empirisch geschätzt.

In Permutationstests ist die Nullhypothese wie folgt definiert: Die Kennzeichnung, die Individuen einer Stichprobe zu Subgruppen (Patienten und Kontrollen) zuordnet sind innerhalb der gesamten Stichprobe austauschbar. Dann deuten signifikante p-Werte darauf hin, dass die Kennzeichnung zu den Subgruppen in der Original-Stichprobe nicht austauschbar, sondern absolut relevant ist (Knijnenburg et al. 2009). Trifft diese Voraussetzung auf eine Stichprobe zu, ist der p-Wert exakt und unverzerrt (Good 2000). Für die Schätzung des p-Werts $\hat{p}$ erfolgt zunächst eine Berechnung der Teststatistik T_O hinsichtlich der Original-Phänotypdaten. Anschließend werden die Phänotypdaten der betrachteten Stichprobe bzgl. aller möglichen Anordnungen permutiert und die zugehörigen Teststatistiken erneut berechnet. Bei der Stichprobengröße n ergeben sich $n!$ mögliche Permutationen des Phänotypvektors.

Schon bei kleinen Fallzahlen ist es aufgrund langer Rechenzeiten kaum möglich, die Teststatistiken aller theoretisch möglichen Permutationen der Phänotypdaten zu berechnen. Da aber jede permutierte Stichprobe Teil der zugrundeliegenden Grundgesamtheit unter der Nullhypothese ist, genügt es nur eine hinreichend große Anzahl von zufällig ausgewählten Permutationen durchzuführen, um einen p-Wert schätzen zu können (Cordell 2009). Diese Vorgehensweise wird als *Monte-Carlo-*

Experiment bezeichnet. Ein Schätzer für den wahren jedoch konservativen p-Wert ist dann:

$$\hat{p} = \frac{1}{N+1}\Big(\sum_{t=1}^{N} I\{T_t \geq T_O\} + 1\Big) \tag{2.10}$$

Damit ist der p-Wert bzgl. der betrachteten Hypothese definiert als der Anteil der permutierten Teststatistiken die mindestens so groß sind wie die Teststatistik auf Basis der Original-Daten (Knijnenburg et al. 2009). Die Erhöhung des Zählers und des Nenners in Gleichung (2.10) um 1 stellt zwar eine Verzerrung des wahren p-Werts dar (die für große N aber vernachlässigbar ist), ist aber aus zwei praktischen Gründen sinnvoll: Zum einen wird 0 als Schätzung ausgeschlossen, zum anderen wird vermieden, dass der wahre p-Wert unterschätzt wird (Phipson und Smyth 2010).

Bei genomweiten Analysen, in denen parallel eine Vielzahl von Hypothesen getestet werden und somit das gesetzte Signifikanzniveau entsprechend adjustiert werden muss, kann trotzdem eine sehr große Anzahl von Permutationen notwendig werden, was einen enormen Rechenaufwand bedeutet (Derkach et al. 2014). Nicht nur die hohe Rechenzeit ist ein Nachteil Permutations-basierter Ansätze, sondern auch die Annahme der Austauschbarkeit unter der Nullhypothese. Goeman und Solari (2014) argumentieren, dass diese Annahme bei genomischen Daten häufig nicht zutrifft.

Madsen und Browning (2009) schlagen für ihre Gruppierungsmethode einen Ansatz vor, der die Rechenzeit drastisch reduziert. Sie gehen davon aus, dass die Teststatistik ihrer Methode (Abschnitt 2.3.6) asymptotisch normalverteilt ist und unbekannten Parametern folgt. Um Mittelwert und Varianz zu schätzen, empfehlen sie, zunächst eine feste Anzahl $B = 1000$ von Permutationen der Phänotypdaten durchzuführen und auf Basis dieser permutierten Daten B Teststatistiken zu berechnen und damit die unbekannten Verteilungsparameter zu schätzen. Der p-Wert wird dann schließlich über die mit den geschätzten Parametern standardisierte Original-Teststatistik aus der Standard-Normalverteilung geschätzt. Sun et al. (2011) stellen diesen Ansatz allerdings in Frage, da nicht klar ist, inwieweit die Permutationsverteilung die Verteilung unter der Nullhypothese widerspiegelt. Weiterhin ist unklar, ob die Annahme der Normalverteilung auch in den Extremen der Permutationsverteilung berechtigt ist, was insbesondere für kleine p-Werte wichtig ist und daher zu verzerrten Fehlern 1. und 2. Art führen kann (Dering et al. 2011). Alternative Ansätze zur

Schätzung der Verteilungsschwänze wurden von Knijnenburg et al. (2009) und Qian (2004) vorgeschlagen. Während erstere eine verallgemeinerte Pareto-Verteilung zur Approximation der Verteilungsenden heranziehen, schätzen letztere die Extreme der Normalverteilung durch eine lineare Extrapolation.

Neben den Permutations-basierten Gruppierungsansätzen gibt es eine Reihe von Methoden, die auf Teststatistiken mit bekannten Verteilung basieren. Dann kann der gesuchte p-Wert leicht über die entsprechende Verteilung geschätzt werden. Bei den meisten Gruppierungsmethoden liegt eine zentral asymptotische χ^2-verteilte Teststatistik vor, insbesondere wenn der zugrundeliegende Phänotyp als normalverteilt angenommen wird. Zu beachten ist allerdings, dass diese Annahme nur gilt, wenn die Varianten einer ROI unabhängig voneinander vererbt wurden und somit nicht im Kopplungsungleichgewicht (engl. linkage disequilibrium, LD) sind und die Stichprobenzahl hinreichend groß ist (Basu und Pan 2011; Zawistowski et al. 2010). Ähnlich dem oben beschriebenen Ansatz von Madsen und Browning (2009) schlagen Lee, Emond et al. (2012) in ihrer Anpassung für kleine Stichprobengrößen eine Schätzung der ersten vier Momente der Verteilung unter der Nullhypothese vor. Ein großer Vorteil solcher Ansätze ist die geringere Rechenzeit, da die Teststatistik anders als bei Permutationsansätzen nur einmal berechnet werden muss. Allerdings liefern χ^2-basierte Teststatistiken in der Situation bei zu kleinen Stichproben zu konservative p-Werte, d.h. der wahre p-Wert wird überschätzt. Dies ist insbesondere bei der Untersuchung seltener Varianten relevant. Ab moderaten Stichprobengrößen liefern χ^2-basierte Teststatistiken häufig zu liberale p-Werte, d.h. der wahre p-Wert wird unterschätzt (Agresti 1992; Mielke und Berry 2007).

2.3 Die untersuchten Gruppierungsmethoden

Da in den vergangenen fünf Jahren weit über 50 Gruppierungsansätze (Auer und Lettre 2015; Basu und Pan 2011; Lee et al. 2014; Moutsianas et al. 2015) vorgeschlagen wurden ist ein Vergleich aller Methoden nicht möglich. Um dennoch die wesentlichen Ideen und Konzepte der bis dato vorgeschlagenen Gruppierungsmethoden darzulegen werden im Folgenden eine entsprechend ausgewählte Menge von insgesamt 15 Gruppierungsmethoden näher betrachtet.

Zu jeder Methode werden die definierenden Eigenschaften genannt und erläutert. Außerdem wird zu jeder Methode ein detaillierter Algorithmus angegeben und auf die Verwandtschaft der einzelnen Ansätze zu anderen hier nicht ausführlich dargestellten Gruppierungsmethoden hingewiesen. Die Eigenschaften aller hier betrachteten Gruppierungsmethoden sind in Adaption zur Tabelle 1 aus einer früheren Arbeit von Dering et al. (2014) in Tabelle 2.1 zusammengefasst.

Tabelle 2.1: Eigenschaften der Gruppierungsmethoden. Autor: Referenz der Veröffentlichung; Jahr: Jahr der Veröffentlichung; Kod.: Kodierung der Varianten der betrachteten Region von Interesse (ROI) als Indikator- (I) oder Genotypkodierung (G); Effektrichtg.: Berücksichtigung verschiedener Effektrichtungen von Varianten in der ROI (J=Ja/N=Nein); Häufig & selten: Gemeinsame Analyse häufiger und seltener Varianten (J/N); Regr.: Ansatz Regressions-basiert (J/N); Phäno.: betrachteter Phänotyp binär (B) oder quantitativ (Q); Kovar.: Einbeziehung von Kovariablen möglich (J/N); Gewichte: Verwendung Varianten-weiser Gewichte (J/N); Struktur der Teststat.: Lineare (L) oder quadratische (Q) Teststatistik nach Derkach et al. (2014); Burden/Non-burden: Burden (B) und Nicht-burden (NB) Methode; Perm.: Schätzung des p-Werts über Permutation (J/N).

Methode	Autor	Jahr	Kod. (I/G)	Effekt-richtg. (J/N)	Häufig & selten (J/N)	Regr. (J/N)	Phäno. (B/Q)	Kovar. (J/N)	Gewichte (J/N)	Struktur der Teststat. (L/Q)	Burden/ Nicht-burden (B/NB)	Perm. (J/N)
CAST	Morgenthaler und Thilly	2007	I	N	N	N	B	N	N	L	B	N
CMC	Li und Leal	2008	I	N	J	N	B	N	N	L	B	J
RVT2	Morris und Zeggini	2010	I	N	N	J	B/Q	J	N	L	B	N
RC	Bhatia et al.	2010	I	N	N	N	B	J	N	L	B	J
WSS	Madsen und Browning	2009	G	N	N	N	B	N	J	L	B	J
RVT1	Morris und Zeggini	2010	G	N	N	J	B/Q	J	N	L	B	N
ASUM	Han und Pan	2010	G	J	N	J	B	N	J	L/Q	B/NB	J
VT	Price et al.	2010	G	J	N	J	B/Q	J	J	L	B	J
KBAC	Liu und Leal	2010	G	N	N	N	B	N	J	L	B	J
CMAT	Zawistowski et al.	2010	G	N	N	N	B	N	N	L	B	J
C-α	Neale et al.	2011	G	J	N	N	B	N	N	Q	NB	N
FPCA	Luo et al.	2011	G	J	N	N	B	J	J	L/Q	B/NB	N
PWST	Zhang et al.	2011	G	J	N	J	B/Q	N	J	L/Q	B/NB	J
SKAT	Wu et al.	2011	G	J	N	J	B/Q	J	J	Q	NB	N
SKAT-O	Lee, Emond et al.	2012	G	J	N	J	B/Q	J	J	Q	NB	N

CAST: Cohort allelic sum test; CMC: Combined multivariate cluster; RVT: Rare variant test 1 und 2; RC: Rarecover; WSS: Weighted sum statistic; ASUM: Adaptive summation; VT: Variable threshold; KBAC: Kernel-based adaptive cluster; CMAT: Cumulative minor-allele test; C-α: C-alpha-based Test; FPCA: Functional principal component analysis; PWST: P-value weighted sum test; SKAT: Sequencing kernel association test; SKAT-O: Optimal unified SKAT

2.3.1 CAST – Cohort Allelic Sum Test

Der Gruppen-basierte Allel-Summierungs-Test (CAST) von Morgenthaler und Thilly (2007) ist einer der ältesten Gruppierungsansätze. Ein ähnlicher Ansatz wurde bereits zuvor in der Arbeit von Fitze et al. (2002) verwendet, aber nicht als eigenständige Methode herausgestellt. In diesem Test geht die ROI über die Indikatorkodierung in den Test ein und ist ausschließlich auf einen dichotomen Phänotypen, typischerweise den Fall-Kontroll-Status, anwendbar. Der Ansatz wurde entwickelt, um ein erhöhtes Vorkommen von seltenen Varianten in einer von zwei betrachteten Gruppen nachzuweisen.

Aufgrund der Indikatorkodierung der ROI wird bei CAST implizit von einer gleichen Effektrichtung sowie Effektstärke der Varianten ausgegangen. Eine Einbeziehung von Kovariablen oder häufigen Varianten ist in CAST nicht vorgesehen. Erweiterungen von CAST, die die Genotypkodierung verwenden oder auch häufige Varianten mit in die Analyse einbeziehen (Abschnitte 2.3.2 und 2.3.7), wurden von Zawistowski et al. (2010) sowie Li und Leal (2008) vorgeschlagen.

Algorithmus

1. Zur Stichprobe $\{1, \ldots, n\}$ seien die beiden Teilmengen der Kontrollen $N^u \subset \{1, \ldots, n\}$ und der Fälle $N^a \subset \{1, \ldots, n\}$ gegeben, so dass für die Anzahlen gilt $\#N^u = n^u$ und $\#N^a = n^a$ mit $n^a + n^u = n$.

2. Zähle die Anzahl der Individuen mit mindestens einem seltenen Allel in der betrachteten ROI, sowohl für die Gruppe der Kontrollen als auch für die Gruppe der Fälle (vgl. Tabelle 2.2).

3. Berechne den p-Wert mit Hilfe Pearsons χ^2-Test oder über den exakten Test nach Fisher (1922), welcher in Gleichung (2.11) dargestellt ist. Die Teststatistik folgt einer hypergeometrischen Verteilung, so dass der p-Wert p^{CAST} wie folgt geschätzt wird:

Tabelle 2.2: 2×2 - Kontingenztabelle: Präsenz bzw. Abwesenheit mindestens eines seltenen Allels in der Region von Interesse bei Fällen und Kontrollen

	Mindestens ein seltenes Allel vorhanden	Kein seltenes Allel vorhanden	Anzahl der Individuen
Fälle	$a = \sum_{i \in N^a} x_i^{\mathrm{I}}$	$b = \sum_{i \in N^a} (1 - x_i^{\mathrm{I}})$	$a + b = n^a$
Kontrollen	$c = \sum_{i \in N^u} x_i^{\mathrm{I}}$	$d = \sum_{i \in N^u} (1 - x_i^{\mathrm{I}})$	$c + d = n^u$
	$a + c$	$b + d$	n

$$p^{\mathrm{CAST}} = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+c)!(a+b)!(c+d)!(b+d)!}{a!b!c!d!n!} \tag{2.11}$$

2.3.2 CMC – Combined and Multivariate Collapsing

Der Kombinationstest aus einer multivariaten und einer Gruppierungsmessgröße (CMC) wurde von Li und Leal (2008) vorgeschlagen und ist eine direkte Weiterentwicklung von CAST, die eine zusätzliche Berücksichtigung von häufigen Varianten erlaubt. Dabei gehen die seltenen Varianten als Gruppe und die häufigen Varianten einzeln in die Berechnung der Teststatistik ein. Die Zusammenfassung der seltenen Varianten erfolgt mittels Indikatorkodierung für die beiden Gruppen Fälle und Kontrollen. Wie auch bei CAST ist die Einbeziehung von Kovariablen für die ursprüngliche CMC Methode nicht vorgesehen. Nach der Klassifizierung von Derkach et al. (2014) (Abschnitt 2.2.5) hat die CMC Methode einer lineare Teststatistik. Die Eigenschaften des CMC-Ansatzes sind in Tabelle 2.1 zusammengefasst.

Algorithmus

1. Entsprechend einer zuvor definierten MAF-Grenze $\tau \leq 5\%$ kombiniere alle Varianten der ROI mit einer MAF $< \tau$ und schätze für diese Gruppe einen p-Wert p^S wie bei CAST in Gleichung (2.11) beschrieben (Abschnitt 2.3.1).

2. Schätze für jede der L häufigen Varianten (MAF $\geq \tau$) der ROI einen p-Wert p_l^H, $l = 1, \ldots, L$ mit Hilfe eines univariaten Tests wie dem χ^2-Test oder dem exakten Test nach Fisher.

3. Kombiniere die erhaltenen p-Werte zu p^{CMC} mit Hilfe einer Kombinationsmethode wie Hotelling's T^2-Test (Hotelling 1992) oder der Kombinationsregel nach Fisher (1992) in Gleichung (2.12):

$$p^{\text{CMC}} = -2\Big(\ln p^S + \sum_{l=1}^{L} \ln p_l^H \Big). \tag{2.12}$$

2.3.3 RC – Rarecover

In dem Ansatz RC von Bhatia et al. (2010) wird nach der optimalen Zusammensetzung der Varianten einer ROI gesucht, wobei die seltenen Varianten (engl. „rare“) durch ein *Fenster* überdeckt (engl. „cover“) werden. Dabei ist die Fenstergröße höchstens so groß wie die gesamte ROI und die optimale Zusammensetzung der Varianten wird über die größte Teststatistik bestimmt.

Mit dieser Gruppierungsmethode ist nur eine Untersuchung binärer Phänotypen ohne die Einbeziehung von Kovariablen möglich. Des Weiteren können keine individuellen Gewichte für die Varianten berücksichtigt werden und die definierten Fenster werden mittels Indikatorkodierung analysiert. RC ist vergleichbar mit dem Ansatz von Price et al. (2010), in dem allerdings die optimale MAF-Grenze gesucht wird, unter welcher alle Varianten als funktionell relevant angenommen werden und somit die Teststatistik maximiert wird (Abschnitt 2.3.8). Die weiteren Eigenschaften der RC-Methode sind in Tabelle 2.1 zusammengefasst.

Es ist zu beachten, dass die Wahl der Variante, die die Teststatistik maximiert, möglicherweise nicht eindeutig ist, da voneinander verschiedene Varianten die gleiche maximale Teststatistik liefern können. In dieser Situation ist nicht klar, welche Variante in die optimale Konstellation einfließen soll. Daher existieren verschiedene Implementationen dieser Methode, die somit für ein und dieselbe ROI unterschiedliche Ergebnisse liefern.

Algorithmus

1. Definiere eine Fenstergröße $W \in \{1, \ldots, K\}$ für die betrachtete ROI mit K Varianten und somit die entsprechenden Basenpaaranfangs- und -endpositionen der zugehörigen Fenster $F_1, \ldots, F_{K-W+1}$. Dabei beginnt F_{k+1} eine Variante weiter rechts als F_k, für alle $k = 1, \ldots, W - K + 1$.

2. Um möglichst exakte p-Werte schätzen zu können und dennoch die Rechenzeit moderat zu halten, definiere eine Grenze Q der verwendeten Teststatistik, ab welcher der p-Wert statt asymptotisch mittels Permutation geschätzt wird.

3. Finde für jedes der definierten Fenster F_k, $k = 1, \ldots, W - K + 1$, die Mengen von Varianten C_k, die die Teststatistik T^{F_k} maximiert, wie folgt:

3.1. Starte mit einer leeren Menge $C_k = \emptyset$ für alle k.

3.2. Führe für das k-te Fenster einen univariaten Test für jede (ROI-)Variante durch und füge diejenige Variante zur Menge C_k hinzu, die den kleinsten p-Wert liefert.

3.3. Füge als nächstes diejenige Variante zur Menge C_k aus dem Fenster F_k hinzu, welche die Teststatistik T^{F_k} weiter erhöht bzw. den p-Wert p^{F_k} entsprechend verringert. Die Schätzung des p-Werts erfolgt dabei wie bei CAST (Abschnitt 2.3.1) über einen Vierfelder-Test, wobei die Variantenmenge C_k hier als ROI zu verstehen ist.

3.4. Wiederhole Punkt **3.3.** solange sich die Teststatistik T^{F_k} erhöht bzw. sich der entsprechende p-Wert p^{F_k} verringert oder alle Varianten des Fensters in C_k enthalten sind.

4. Berechne die Teststatistik für die betrachtete ROI als Optimum aller berechneten Teststatistiken der jeweiligen Fenster:

$$T^{\mathrm{RC}} = \max_{k \in \{1,\ldots,W-K+1\}} T^{F_k} \tag{2.13}$$

5. Definiere eine Anzahl von evtl. durchzuführenden Permutationen L um den p-Wert möglichst genau schätzen zu können. Falls $T^{\mathrm{RC}} > Q$ permutiere den Fall-Kontroll-Status in der Stichprobe L mal und führe für jede der permutierten Daten, die Punkte **3.1.** bis **3.4.** in dem betreffenden Fenster erneut durch, so dass Teststatistiken $T_1^{\mathrm{RC}}, \ldots, T_L^{\mathrm{RC}}$ resultieren.

6. Berechne den p-Wert p^{RC} als Anteil der permutierten Teststatistiken T_ℓ^{RC}, $\ell \in \{1, \ldots, L\}$, die mindestens so groß sind wie die Original-Statistik T^{RC} aufweisen,

$$p^{\mathrm{RC}} = \frac{1}{L+1}\Big(\sum_{\ell=1}^{L} I\{T_\ell^{\mathrm{RC}} \geq T^{\mathrm{RC}}\} + 1\Big).$$

2.3.4 RVT1 – Rare Variant Test 1

Der Seltene-Varianten-Test 1 (RVT1) ist der erste von zwei von Morris und Zeggini (2010) vorgeschlagenen Ansätzen zur Untersuchung der Assoziation seltener Varianten mit einem Phänotyp. Diese Methode ist ein Regressions-Ansatz, in dem die Genotypkodierung genutzt wird, wobei die Genotypen aber vorverarbeitet in das Modell einfließen. Auf diese Weise geht der Anteil der Varianten, die mindestens ein seltenes Allel bzgl. der betrachteten ROI haben, als erklärende Variable in das Modell ein. Es wird daher bei diesem Ansatz ein gemeinsamer Effekt der Varianten der ROI angenommen.

Die Berücksichtigung von Kovariablen ist im Allgemeinen vorgesehen, ebenso wie die Betrachtung quantitativer und qualitativer Phänotypen. Da die erklärende Variable über die die Anzahl K der Varianten der ROI gemittelt wurde und auch keine individuellen Gewichte für die Varianten Verwendung finden, werden mögliche bi-direktionale Effekte verwischt. Als eine Erweiterung zu RVT1 schlugen Asimit und Zeggini (2009) den Einbau von Gewichten bzgl. der individuellen Sequenzierungsqualität für jede der Varianten in der ROI vor, um die Zuverlässigkeit der Ergebnisse zu erhöhen. Der positive Effekt dieser Vorgehensweise wurde bisher jedoch noch nicht weiter untersucht. Für weitere Eigenschaften von RVT1 wird auf Tabelle 2.1 verwiesen.

In der folgenden Beschreibung des Algorithmus wird nur das Regressionsmodell für die Betrachtung quantitativer Phänotypen beschrieben. Das Modell für qualitative Phänotypen besitzt eine analoge Form und wird mit Hilfe der logit-Funktion wie in Gleichung (2.4) gebildet.

Algorithmus

1. Bestimme den Anteil der Varianten innerhalb der ROI , die mindestens ein seltenes Allel tragen:

$$\frac{1}{K}\sum_{j=1}^{K} x_{ij}^{\mathrm{G}}$$

2. Das Regressionsmodell hat für jedes Individuum $i \in \{1,\ldots,n\}$ auf Basis von Gleichung (2.3) die Gleichung:

$$y_i = \beta_0 + \beta\Big(\frac{1}{K}\sum_{j=1}^{K} x_{ij}^{\mathrm{G}}\Big) + \boldsymbol{\alpha}^{\intercal}\boldsymbol{c}_i + \varepsilon_i. \tag{2.14}$$

Dabei ist y_i der quantitative Phänotyp für Person i, β_0 ist der Achsenabschnitt und der Effektvektor $\boldsymbol{\beta}$ aus Gleichung (2.3) reduziert sich zu einem kollektiven Effekt, so dass $\boldsymbol{\beta} = (\beta,\ldots,\beta)^{\intercal}$ gilt. Für die Gewichte aus Gleichung (2.3) gilt $w_j = 1/K$ für alle $j = 1,\ldots,K$, $\boldsymbol{c_i}$ bezeichnet den Vektor der Kovariablen mit Koeffizienten $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_Q)$ und ε_i ist der Fehlerterm. Unter der

Voraussetzung, dass von zwei Individuen eines mindestens ein seltenes Allel in der ROI aufweist und das andere nicht, kann β als Anstieg in der Ausprägung des (quantitativen) Phänotyps interpretiert werden.

3. Schätze die Parameter β_0, β und $\boldsymbol{\alpha}$ des Modells.

4. Teste die Nullhypothese $H_0 : \beta = 0$ mittels eines Score-, Likelihood-Quotienten- oder Wald-Tests, die asymptotisch äquivalent sind (Bera und Bilias 2001; Engle 1984).

2.3.5 RVT2 – Rare Variant Test 2

Der Seltene-Varianten-Test 2 (RVT2), ebenfalls von Morris und Zeggini (2010), unterscheidet sich von RVT1 durch die Verwendung der Indikator- statt Genotypkodierung für die Varianten innerhalb der Region von Interesse. Das heißt für jedes Individuum der Stichprobe wird nur geprüft ob (innerhalb der ROI) mindestens ein seltenes Allel existiert oder nicht. Wie im Ansatz von RVT1 wird die Information bzgl. der ROI auf eine Art Super-Variante projiziert, so dass keine separaten Gewichte für jede Variante verwendet werden können und ein gemeinsamer Effekt aller Varianten angenommen wird. Demnach können auch hier bi-direktionale Effekte nicht erkannt werden. Die Assoziation zum betrachteten Phänotyp wird mit der Indikator-kodierten ROI als unabhängiger Variable mittels Regression und einem geeigneten statistischen Test geprüft. Dadurch ist der Ansatz hinsichtlich der Art des Phänotyps flexibel und ermöglicht die Verwendung von Kovariablen. Die Eigenschaften von RVT2 sind in Tabelle 2.1 nochmals zusammengefasst.

Algorithmus

1. Bestimme die x_i^{I} für alle Individuen $i \in \{1, \ldots, n\}$ der Stichprobe.

2. Das Regressionsodell auf Basis von Gleichung (2.3) nimmt dann die Form an:

$$y_i = \beta_0 + \beta x_i^I + \boldsymbol{\alpha}^\intercal \boldsymbol{c}_i + \varepsilon_i. \tag{2.15}$$

   Dabei ist y_i der quantitative Phänotyp für Person i, β_0 ist der Achsenabschnitt und der Effektvektor $\boldsymbol{\beta}$ aus Gleichung (2.3) reduziert sich zu einem kollektiven Effekt β der auf die gesamte ROI wirkt. Für die Gewichte aus Gleichung (2.3) gilt $w_j = 1$ für $j = 1, \ldots, K$, $\boldsymbol{c_i}$ bezeichnet den Vektor der Kovariablen mit Koeffizienten $\boldsymbol{\alpha}$ und ε_i ist der Fehlerterm. Dabei wird β als die quantitative Veränderung der Ausprägung des Phänotyps betrachtet zwischen einer Person, die mindestens ein seltenes Allel in der ROI trägt im Vergleich zu einer Person, die keins trägt.

3. Schätze die Parameter β_0, β und $\boldsymbol{\alpha}$.

4. Prüfe die Nullhypothese $H_0 : \beta = 0$ mittels eines Score-, Likelihood-Quotienten- oder Wald-Tests (Bera und Bilias 2001; Engle 1984).

2.3.6 WSS – Weighted Sum Statistic

Die gewichtete Summierungs-Methode (WSS) von Madsen und Browning (2009) war die erste Gruppierungsmethode, die individuelle Gewichte für die verschiedenen seltenen Varianten einer ROI nutzte. Dabei wird für jede Variante j, $j \in \{1, \ldots, K\}$ und jedes Individuum i, $i \in \{1, \ldots, n\}$ die Anzahl x_{ij} von seltenen Allelen durch die entsprechende Abweichung in der Gruppe der Kontrollen normiert. Die Idee zur Verwendung dieser Gewichte basiert auf der Tatsache, dass die Effektstärke von seltenen Varianten umgekehrt proportional zu ihrer Allelfrequenz ist Manolio et al. (2009).

Der originale Ansatz von Madsen und Browning ist ausschließlich zur Untersuchung von seltenen Varianten einsetzbar und sieht keine Berücksichtigung von Kovariablen vor. Die zugehörige Teststatistik gehört nach der Definition von Derkach et al. (2014) zur Klasse der linearen Teststatistiken. Für weitere Eigenschaften von WSS sei auf Tabelle 2.1 verwiesen.

Erweiterungen der Methode wurden von Feng et al. (2011) und Lin und Tang (2011) vorgeschlagen. Während Lin und Tang in ihrer Methode auch die Betrachtung quantitativer Phänotypen und die Einbeziehung von Kovariablen erlauben, schlagen Feng et al. Gewichte auf Basis des Odds-Ratios einer Variante anstelle der geschätzten Standardabweichung vor. Im Folgenden wird der Algorithmus für den Originalansatz beschrieben.

Algorithmus

1. Zur Stichprobe $\{1, \ldots, n\}$ seien die beiden Teilmengen der Kontrollen $N^u \subset \{1, \ldots, n\}$ und der Fälle $N^a \subset \{1, \ldots, n\}$ gegeben, so dass für die Anzahlen gilt $\#N^u = n^u$ und $\#N^a = n^a$ mit $n^a + n^u = n$.

2. Schätze die Frequenz des seltenen Allels von Variante j in der Kontrollgruppe der betrachteten Stichprobe für alle $j \in \{1, \ldots, K\}$ wie folgt:

$$\hat{p}_j^u = \frac{\sum_{i \in N^u} x_{ij}^G + 1}{2(n^u + 1)}. \tag{2.16}$$

Dabei entspricht x_{ij}^{G} dem Genotyp von Person i in der Variante j wie in Gleichung (2.1) beschrieben.

3. Schätze die Varianten-basierte Gewichte:

$$w_j^{\mathrm{WS}} = \frac{1}{\sqrt{n\hat{p}_j^u(1 - \hat{p}_j^u)}}. \tag{2.17}$$

4. Berechne für jedes Individuum $i \in \{1, \ldots, n\}$ der Stichprobe den sogenannten genetischen Score:

$$x_i^{\mathrm{WS}} = \sum_{j=1}^{K} w_j^{\mathrm{WS}} x_{ij}^{\mathrm{G}} \tag{2.18}$$

5. Sortiere die genetischen Scores der gesamten Stichprobe der Größe nach aufsteigend und vergib entsprechende Ränge: $r_i = \text{Rang}(x_i^{\text{WS}})$, $i = 1, \ldots, n$.

6. Berechne die Teststatistik der WSS-Methode aus der Rangsumme aller genetischen Scores der Kontrollgruppe:

$$T^{\text{WS}} = \sum_{i \in N^u} r_i \tag{2.19}$$

7. Permutiere den Fall-Kontroll-Status L mal und führe die Schritte Punkte **2.** bis **6.** für jede Permutation erneut durch, so dass sich die Rangsummen-Statistiken T_ℓ^{WS}, $\ell \in \{1, \ldots, L\}$ ergeben.

8. Der p-Wert p^{WS} kann auf zwei Arten berechnet werden:

In der klassischen Variante ist der p-Wert definiert als Anteil der permutierten Rangsummen-Statistiken T_ℓ^{WS}, $\ell \in \{1, \ldots, L\}$, die mindestens so groß sind wie die Original-Rangsummen-Statistik T^{WS}.

$$p^{\text{WS}} = \frac{1}{L+1}\Big(\sum_{\ell=1}^{L} I\{T_\ell^{\text{WS}} \geq T^{\text{WS}}\} + 1\Big)$$

Die zweite Variante bietet, wie bereits in Abschnitt 2.2.6 erwähnt, eine erhebliche Ersparnis an Rechenzeit. Dabei werden der Mittelwert ($\hat{\mu}_B$) und die Standardabweichung ($\hat{\sigma}_B$) der empirischen Verteilungsfunktion unter der Nullhypothese nach Durchführung von $B \ll L$ Permutationen geschätzt. Madsen und Browning (2009) schlagen $B = 1000$ vor. Dabei wird angenommen, dass die empirische Verteilung approximativ einer Normalverteilung mit Mittelwert $\hat{\mu}_B$ und Varianz $\hat{\sigma}_B^2$ folgt. Dann wird der p-Wert p^{WS} mittels der Standard-Normalverteilung der standardisierten Rangsummen-Statistik $(T^{\text{WS}} - \hat{\mu}_B)/\hat{\sigma}_B$ geschätzt.

2.3.7 CMAT – Cumulative Minor–Allele Test

Der Kumulierte-Seltene-Allele-Test (CMAT) von Zawistowski et al. (2010) kann als genotypbasierte Variante von CAST verstanden werden (Morgenthaler und Thilly (2007), Abschnitt 2.3.1). Im Gegensatz zu CAST werden bei CMAT nicht die Fälle und Kontrollen hinsichtlich der Existenz mindestens eines Risikoallels einander gegenübergestellt, sondern die jeweilige Anzahl der Risiko-Allele und der Wildtyp-Allele in den Fällen und den Kontrollen verglichen. Dabei ist das Wildtyp-Allel dasjenige, was für gewöhnlich in Natura bzw. in der betrachteten Population zu erwarten ist.

Der originale CMAT Ansatz ist ein einfacher Gruppierungsansatz für seltene Varianten und bindet weder Gewichte für einzelne Varianten noch häufige Varianten in die Untersuchung ein, wie in Tabelle 2.1 zusammengefasst ist. Aufgrund der Struktur der Teststatistik von CMAT ist der Ansatz nicht robust in Gegenwart von bi-direktionalen Effekten in der ROI. Des Weiteren gibt es in der ursprünglichen Methode nicht die Möglichkeit Kovariablen zu berücksichtigen. Allerdings schlagen Zawistowski et al. (2010) eine Erweiterung ihres Ansatzes um die Berücksichtigung von kategoriellen Kovariablen vor. Dabei werden die Anzahlen der Risiko- und Wildtyp-Allele einzeln für jede in der Stichprobe präsente Kategorie der Kovariablen in den Fällen und Kontrollen betrachtet und zu einer Teststatistik aufsummiert. Da für die statistische Untersuchung nur eine Implementation ohne Kovariablen zur Verfügung steht, wird hier ausschließlich der Original-Ansatz vorgestellt.

Algorithmus

1. Zur Stichprobe $\{1, \ldots, n\}$ seien die beiden Teilmengen der Kontrollen $N^u \subset \{1, \ldots, n\}$ und der Fälle $N^a \subset \{1, \ldots, n\}$ gegeben, so dass für die Anzahlen gilt $\#N^u = n^u$ und $\#N^a = n^a$ mit $n^a + n^u = n$.

2. Zähle die Anzahl von Risiko- und Wildtyp-Allelen auf Basis der zugrundeliegenden Genotypen in der ROI aus Gleichung (2.1) sowohl in der Gruppe der Fälle als auch in den Kontrollen wie in Tabelle 2.3 angegeben.

Tabelle 2.3: 2×2-Kontingenztabelle: Anzahl der Risiko- und Wildtyp-Allele bei Fällen und Kontrollen bzgl. einer Region von Interesse.

	Risiko-Allele	Wildtyp-Allele	
Fälle	$a = \sum_{i \in N^a} \sum_{j=1}^{K} x_{ij}$	$b = \sum_{i \in N^a} \sum_{j=1}^{K} (2 - x_{ij})$	$a + b = 2Kn^a$
Kontrollen	$c = \sum_{i \in N^u} \sum_{j=1}^{K} x_{ij}$	$d = \sum_{i \in N^u} \sum_{j=1}^{K} (2 - x_{ij})$	$c + d = 2Kn^u$
	$a + c$	$b + d$	$2Kn$

3. Berechne die Teststatistik gemäß:

$$T^{\text{CMAT}} = \frac{n}{2Kn^a n^u} \times \frac{(ad - bc)^2}{(a + c)(b + d)}$$

4. Permutiere die Stichprobe bzgl. des Fall-Kontrollstatus L mal und führe die Schritte **2.** und **3.** erneut auf den permutierten Daten durch. Berechne den p-Wert:

$$p^{\text{CMAT}} = \frac{1}{L + 1} \Big(\sum_{\ell=1}^{L} I\{T_\ell^{\text{CMAT}} \geq T^{\text{CMAT}}\} + 1 \Big)$$

Dabei sind $T_\ell^{\text{CMAT}}, \ell = 1, \ldots, L$, die Teststatistiken und I ist die Indikatorfunktion, die den Wert 1 annimmt, falls die Bedingung in der geschweiften Klammer erfüllt ist, und 0, falls nicht.

2.3.8 VT – Variable Threshold

In der Methode des variablen Schwellenwerts (VT) von Price et al. (2010) wird nach der optimalen Grenze für die Frequenz der seltenen Allele der Varianten einer ROI gesucht. VT ist ähnlich zur RC-Methode von Bhatia et al. (2010) (Abschnitt 2.3.3) mit dem Unterschied, dass nicht nach der optimal zusammengesetzten Teilmenge von Varianten einer ROI, sondern nach einer optimalen MAF-Grenze gesucht wird. VT liegt der Annahme zugrunde, dass die Varianten unter einer bestimmten MAF-Grenze eine höhere Wahrscheinlichkeit haben, funktionell relevant zu sein. Die Idee der Methode basiert auf mehreren Beobachtungen. Zum einen ist der Anteil der seltenen Varianten innerhalb des menschlichen Genoms sehr viel größer als der

Anteil der häufigen Varianten (Burkett und Greenwood 2013). Des Weiteren ist die Effektstärke der seltenen oft sehr viel größer als die häufiger Varianten, woraus die Annahme eines Zusammenhangs zwischen der funktionellen Relevanz und der Frequenz des seltenen Allels einer Variante entstanden ist (Manolio 2010; Price et al. 2010). Allerdings geht eine Erhöhung der MAF-Grenze auch mit einer Erhöhung der Varianz einher, was ein Nachteil der VT-Methode ist (Dering et al. 2011). Daher muss bei der Suche nach der optimalen MAF-Grenze zwischen dem potentiellen Informationsgewinn und der steigenden Varianz abgewogen werden.

Eine der Hauptannahmen der VT-Methode ist, dass die Varianten der betrachteten ROI den Phänotyp nur in eine Richtung beeinflussen. Für jede Person der Stichprobe wird der Genotyp jeder Variante mit der individuellen Abweichung der beobachteten Phänotyp-Ausprägung der Person zum Stichprobenmittelwert gewichtet. Da in den meisten Analysen zunächst nicht klar ist, ob die ROI Varianten mit unterschiedlichen Effektrichtungen besitzt, kann dies Auswirkungen auf die Teststärke haben, da sich vorhandene Effekte ggf. auslöschen.

In einer Erweiterung ihres Ansatzes schlagen die Autoren vor, Vorhersagewerte für die funktionelle Schädlichkeit der einzelnen Varianten als Gewichte in die Teststatistik einzubringen, was hier nicht untersucht wird. Der VT-Ansatz wurde sowohl für die Untersuchung von quantitativen als auch dichotomen Phänotypen sowie in einer Erweiterung für die Einbeziehung eventueller Kovariablen vorgeschlagen (vgl. Tabelle 2.1). Da der Ansatz sehr rechenintensiv ist und nur eine Implementation ohne Verwendung von Kovariablen zur Verfügung, muss in dieser Arbeit auf die Einbeziehung von Kovariablen verzichtet werden und nur der quantitative Phänotyp wird betrachtet.

Algorithmus

1. Zur Stichprobe $\{1, \ldots, n\}$ seien die beiden Teilmengen der Kontrollen $N^u \subset \{1, \ldots, n\}$ und der Fälle $N^a \subset \{1, \ldots, n\}$ gegeben, so dass für die Anzahlen gilt $\#N^u = n^u$ und $\#N^a = n^a$ mit $n^a + n^u = n$.

2. Berechne für jede Variante $j \in \{1, \ldots, K\}$ der ROI die Frequenz maf_j des seltenen Allels:

$$\text{maf}_j = \frac{1}{2n} \sum_{i=1}^{n} x_{ij} \tag{2.20}$$

Dabei ist x_{ij} der Genotyp von Variante j in Individuum i wie in Gleichung (2.1).

3. Ordne die berechneten MAFs maf_j, $j \in \{1, \ldots, K\}$, der Größe nach und reduziere mehrfach vorkommende MAF-Grenzen auf jeweils eine Grenze, so dass K', $K' \leq K$, eindeutige MAF-Grenzen $\tau_1, \ldots, \tau_{K'}$ bezüglich der betrachteten ROI entstehen.

4. Berechne zu jeder MAF-Grenze τ_k die Teststatistik $T^{\text{VT}}(\tau_k)$ gemäß:

$$T^{\tau_k} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{K} w_j^{\tau_k} x_{ij} (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{K} (w_j^{\tau_k} x_{ij})^2}}, \quad k = 1, \ldots, K', \tag{2.21}$$

mit der quantitativen Phänotypausprägung y_i und dem Mittelwert $\overline{y}$ der Phänotyp-Ausprägung bzgl. der Stichprobe. Die Gewichte $w_j^{\tau_k}$ der Varianten $j = 1, \ldots, K$ sind über die Indikatorfunktion definiert:

$$w_j^{\tau_k} = I\{\text{maf}_j \leq \tau_k\} \tag{2.22}$$

5. Bestimme die maximale Teststatistik über alle MAF-Grenzen $\tau_1, \ldots, \tau_{K'}$:

$$T^{\text{VT}} = \max_{k \in \{1, \ldots, K'\}} T^{\tau_k} \tag{2.23}$$

6. Permutiere die Stichprobe bzgl. des Phänotyps L mal und führe die Punkte **4.** und **5.** erneut auf den permutierten Daten durch. Berechne den p-Wert:

$$p^{\text{VT}} = \frac{1}{L+1} \Big(\sum_{\ell=1}^{L} I\{T_\ell^{\text{VT}} \geq T^{\text{VT}}\} + 1 \Big).$$

2.3.9 ASUM – Adaptive Summation Test

Die adaptive Summierungsmethode (ASUM) von Han und Pan (2010) ist eine der ersten, die bi-direktionale Effekte von Varianten innerhalb einer ROI berücksichtigt. Der Kern der ASUM-Methode ist eine Regression, in die die Genotypkodierung einfließt. Kovariablen werden nicht berücksichtigt. Die Autoren schlagen Daten-adaptive Gewichte vor, die als Koeffizienten eines multiplen Regressionsmodells für jede der Varianten der ROI geschätzt werden, um eventuelle bi-direktionale Effekte zu berücksichtigen.

Die ursprüngliche Methode zielte nur auf die Untersuchung seltener Varianten und binärer Phänotypen (vgl. Tabelle 2.1). Hoffmann et al. (2010) entwickelten den Ansatz für quantitative Phänotypen weiter. Daneben stellten Han und Pan (2010) in ihrer Arbeit einen ähnlichen Ansatz wie Li und Leal (2008) vor, um häufige Varianten in die Analyse einzubeziehen. Die erweiterte Methode wird mit aSumC bezeichnet. Im Folgenden wird nur der ursprüngliche Ansatz ASUM vorgestellt.

Da ASUM ein Daten-adaptiver Test ist, muss der p-Wert über Permutation geschätzt werden. Um Rechenzeit zu sparen, schlagen Han und Pan (2010) einen ähnlichen Ansatz wie Madsen und Browning (2009) vor. Hierbei wird zunächst nur eine kleine Anzahl von Permutationen betrachtet mit deren Hilfe die ersten zwei Momente, Mittelwert und Varianz der Verteilung unter Nullhypothese geschätzt werden. Schließlich kann der p-Wert über die geschätzte Verteilung unter der Nullhypothese und die originale Teststatistik geschätzt werden. Han und Pan (2010) weisen aber darauf hin, dass dieser Ansatz möglicherweise nicht immer zuverlässige Werte liefert. Daher wird in dieser Arbeit nur der originale Permutationsansatz verwendet.

Algorithmus

1. Schätze für jede Variante $j \in \{1, \ldots, K\}$ der ROI die marginalen Effekte $\tilde{\beta}_j$ und den marginalen Achsenabschnitt $\tilde{\beta}_j^0$ mittels des Regressionsmodells:

$$\operatorname{logit} P(y_i = 1) = \tilde{\beta}_{0j} + \tilde{\beta}_j^{\intercal} x_{ij} \tag{2.24}$$

Dabei ist y_i der Fall-Kontroll-Status für Individuum i, $\tilde{\beta}_{0j}$ ist der Achsenabschnitt, $\tilde{\beta}_j$ ist die Effektstärke der Varianten j der ROI und x_{ij} ist der Genotyp der Variante j für Person i, entsprechend der Kodierung aus Gleichung (2.1).

2. Teste für jede Variante j der ROI die Nullhypothese $H_0 : \tilde{\beta}_j = 0$, dass kein Effekt vorliegt, mit resultierenden p-Werten $\tilde{p}_j$.

3. Prüfe für jede Variante j der ROI ob der $\tilde{p}_j$ aus Punkt **2** kleiner ist als ein zuvor definierter Schwellwert, τ^{ASUM} und der zugehörige geschätzte Regressions-Koeffizient $\hat{\tilde{\beta}}_j < 0$ ist, da dies ein Indiz dafür wäre, dass die jeweilige Variante einen protektiven Effekt hat. Die Autoren Han und Pan (2010) empfehlen $\tau^{\text{ASUM}} = 0{,}1$. Aktualisiere dementsprechend die Kodierung der Variante $j \in \{1, \ldots, K\}$ für Person $i \in \{1, \ldots, n\}$ wie folgt:

$$x_{ij}^* = \begin{cases} 2 - x_{ij} & \hat{\tilde{\beta}}_j < 0 \text{ und } \tilde{p}_j < \tau^{\text{ASUM}} \\ x_{ij} & \text{sonst} \end{cases} \tag{2.25}$$

Diese Umkodierung hat zur Folge, dass protektive seltene Varianten für die Schätzung des Effekts nicht berücksichtigt werden, falls eine zweifache Kopie des seltenen Allels vorliegt, und somit er Effekt der von schädigenden Varianten ausgeht mehr Gewicht erhält.

4. Nutze die aktualisierte Kodierung aus Gleichung (2.25), um durch ein neues Regressionsmodell den gemeinsamen Effekt β für die gesamte ROI zu schätzen:

$$\operatorname{logit} P(y_i = 1) = \beta_0 + \beta \sum_{j=1}^{K} x_{ij}^*. \tag{2.26}$$

5. Berechne auf Basis des Modells aus Gleichung (2.26) eine Score- oder Wald-Teststatistik T^{ASUM} (Bera und Bilias 2001; Engle 1984). Da die Kodierung der Varianten Daten-spezifisch geändert wird, ist die Berechnung des p-Werts über die bekannte Verteilung nicht zulässig.

6. Permutiere den Fall-Kontroll-Status L mal und wiederhole die Punkte **1.** bis **5.** für alle permutierten Daten, so dass die Teststatistiken T_ℓ^{ASUM}, $\ell = 1, \ldots, L$, resultieren.

7. Berechne den p-Wert als Anteil der Teststatistiken T_ℓ^{ASUM}, $\ell = 1, \ldots, L$, die mindestens den Wert der originalen Teststatistik T^{ASUM} erreichen.

$$p^{\text{ASUM}} = \frac{1}{L+1}\Big(\sum_{\ell=1}^{L} I\{T_\ell^{\text{ASUM}} \geq T^{\text{ASUM}}\} + 1\Big).$$

2.3.10 KBAC – Kernel Based Adaptive Cluster

Im Ansatz von Liu und Leal (2010) wurde die Untersuchung von seltenen Varianten mit der Methode der Kern-basierten adaptiven Gruppen (KBAC) vorgeschlagen, in der unterschiedliche Genotyp-Strukturen bzgl. einer ROI in einer Stichprobe untersucht werden. Bei der Methode KBAC wird die Gruppierung in zweierlei Hinsicht umgesetzt. Zunächst werden, wie bei den meisten Gruppierungsmethoden, Varianten einer ROI gruppiert. Anschließend werden Gruppen von verschiedenen Genotyp-Vektoren der ROI adaptiv zur vorliegenden Stichprobe gebildet. Für die gefundenen Genotyp-Vektoren wird auf Basis ihrer Häufigkeit in der Gesamtstichprobe und in der Gruppe der Fälle ein relatives Risiko geschätzt. Dies wird dann in einer Teststatistik als Gewicht für die relativen Häufigkeiten in den Fällen und den Kontrollen verwendet.

Die KBAC-Methode ist aufgrund ihrer Konstruktion robust gegenüber bi-direktionalen Effekten, sofern keine neutralen Varianten in der ROI enthalten sind. Durch die Präsenz vereinzelter neutraler Varianten innerhalb der Stichprobe kann eine andere Genotypstruktur als die eigentlich kausale als kausal bestimmt werden oder durch das Vorhandensein neutraler Varianten werden eine Vielzahl verschiedener Genotyp-Vektoren erzeugt, so dass die eigentlich kausalen Varianten nicht genügend Gewicht bekommen, und der vorhandene Effekt verwischt wird. Eine Zusammenfassung der Eigenschaften von KBAC findet sich in Tabelle 2.1.

Die ursprüngliche Methode sieht nur die Untersuchung von dichotomen Phänotypen vor und ist nicht für die Adjustierung von Kovariablen ausgelegt. In ihrer Arbeit

geben Liu und Leal (2010) einen Ansatz, der die Möglichkeit zur Stratifikation nach weiteren Einflussvariablen ermöglicht. Dabei werden die Gewichte und die Kovariablen in ein logistisches Regressionsmodell eingebettet. Auch die gemeinsame Betrachtung seltener und häufiger Varianten ist über diesen erweiterten Ansatz möglich. Da dieser sehr rechenintensiv ist, geben die Autoren für große Stichprobengrößen eine Monte-Carlo-Verfahren an, das mit einer geringeren Rechenzeit auskommt als die ursprüngliche Methode (Liu und Leal 2010). In dieser Arbeit wird der KBAC-Ansatz ohne Kovariablen vorgestellt.

Die Idee, die Struktur der ROI einer gegebenen Stichprobe zu betrachten, wird auch im C-α-Ansatz von Neale et al. (2011) verfolgt. Allerdings vergleichen die Autoren hier die Schwankungen der Anzahl der seltenen Allele in den Fällen im Vergleich zur gesamten Stichprobe (vgl. Abschnitt 2.3.11).

Algorithmus

1. Zur Stichprobe $\{1, \ldots, n\}$ seien die beiden Teilmengen der Kontrollen $N^u \subset \{1, \ldots, n\}$ und der Fälle $N^a \subset \{1, \ldots, n\}$ gegeben, so dass für die Anzahlen gilt $\#N^u = n^u$ und $\#N^a = n^a$ mit $n^a + n^u = n$.

2. Für jede Person $i \in \{1, \ldots, n\}$ der Stichprobe gibt es für die Varianten der betrachteten ROI eine Folge von K Genotypen, die in einem Vektor $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{iK})^\intercal$ zusammengefasst werden. Es seien daher $\{\boldsymbol{G}_1, \boldsymbol{G}_2, \ldots, \boldsymbol{G}_P\} \subset \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ die Menge der eindeutigen Genotyp-Vektoren der gesamten Stichprobe.

3. Dann kann das empirische Stichprobenrisiko für eine Person mit dem Genotyp-Vektor $\boldsymbol{G_k} \in \{\boldsymbol{G}_1, \boldsymbol{G}_2, \ldots, \boldsymbol{G}_P\}$ zu erkranken geschätzt werden als:

$$\hat{R}_k = \frac{n_k^a}{n_k}, \quad k = 1, \ldots, P. \tag{2.27}$$

4. Schätze für jeden Genotyp-Vektor $\boldsymbol{G}_k$, $k = 1, \ldots, P$, ein Gewicht w_k^{KBAC}, das definiert ist als Wahrscheinlichkeit, ein Stichprobenrisiko von höchstens $\hat{R}_k$ zu erreichen:

$$w_k^{\mathrm{KBAC}} = \int_0^{\hat{R}_k} \varphi_k^0(r)dr \tag{2.28}$$

Die Kernfunktion φ_k^0 kann abhängig von der Stichprobengröße verschiedene Formen haben, die die Wirklichkeit am besten wiedergeben. Für kleine Stichprobengrößen empfehlen die Autoren, die hypergeometrische Verteilung zu wählen. Die Form des hypergeometrischen Kern $\varphi_k^0(r_k)$ ist:

$$\varphi_k^0(r_k) = \frac{\binom{n_k}{n_k r_k}\binom{n-n_k}{n^a-n_k r_k}}{\binom{n}{n^a}} \tag{2.29}$$

Da die hypergeometrische Verteilung diskret ist, reduziert sich das Integral in (2.28) zu einer Summe:

$$w_k^{\mathrm{KBAC}} = \sum_{r_k \in \{0/n_k, \ldots, \hat{R}_k\}} \varphi_k^0(r_k) \tag{2.30}$$

5. Berechne die Teststatistik für die Stichprobe gemäß:

$$T^{\mathrm{KBAC}} = \left(\sum_{k=1}^{P} w_k^{\mathrm{KBAC}} \left(\frac{n_k^a}{n^a} - \frac{n_k^u}{n^u} \right) \right)^2 \tag{2.31}$$

T^{KBAC} vergleicht die relativen Häufigkeiten der verschiedenen Genotyp-Vektoren in den Fällen und Kontrollen und gewichtet diese mit den geschätzten Gewichten w_k^{KBAC} bzgl. des Stichprobenrisikos.

6. Permutiere den Fall-Kontrollstatus der Stichprobe L mal und führe die Punkte **3.** bis **5.** erneut auf den permutierten Daten durch. Schätze den p-Wert p^{KBAC} als Anteil der permutierten Teststatistiken $T_\ell^{\mathrm{KBAC}}, \ell = 1, \ldots, L$, die mindestens so groß sind wie die Original-Teststatistik T^{KBAC} haben,

$$p^{\mathrm{KBAC}} = \frac{1}{L+1}\Big(\sum_{\ell=1}^{L} I\{T_\ell^{\mathrm{KBAC}} \geq T^{\mathrm{KBAC}}\} + 1 \Big).$$

2.3.11 C-α – C-alpha-based Test

Die C-alpha-basierte Methode C-α wurde von Neale et al. (2011) auf der Grundlage von zwei früheren Arbeiten entwickelt (Neyman und Scott 1966; Zelterman und Chen 1988). Der Ansatz basiert auf einer variantenweisen Untersuchung einer extremen Abweichung der beobachteten zur erwarteten Anzahl der seltenen Allele in den Fällen im Vergleich zur Abweichung in der Gesamtstichprobe. Vereinfacht ausgedrückt vergleichen Neale et al. (2011) die Verteilung von seltenen Varianten innerhalb einer Stichprobe mit dem Werfen einer fairen Münze: Falls eine Variante nicht krankheitsverursachend bzw. neutral ist, ist die Wahrscheinlichkeit für ihr Vorkommen in der Kontrollgruppe und in der Fallgruppe gleich groß, also 0,5. Das Auftreten einer schädigenden oder protektiven Variante bei der die Wahrscheinlichkeiten, dass sie in der Gruppe der Fälle oder der Kontrollen auftaucht, ungleich 0,5 sind, ist gleichbedeutend mit dem Werfen einer unfairen Münze.

Die C-α-Methode basiert auf dieser einfachen Idee, wobei der Fall-Kontrollstatus ohne Einbeziehung von Kovariablen betrachtet wird (vgl. Tabelle 2.1). Der Ansatz ist in der Lage, sowohl mit neutralen als auch mit bi-direktionalen Effekten innerhalb einer ROI umzugehen. So erhöhen neutrale Varianten die Teststatistik auf Grund der in der Theorie nicht vorhandenen Abweichung nicht, während sowohl protektive als auch schädigende Varianten den Wert der Teststatistik proportional zur Abweichung in der Fallgruppe bzw. der Gesamtstichprobe erhöhen. Ein Nachteil der C-α-Methode ist, dass sogenannte privaten Varianten, die lediglich einmal in der Stichprobe vorkommen, nur ein geringes Gewicht zukommt, obwohl sie krankheitsverursachend sein können. Um dieses Problem zu umgehen, schlagen Neale et al. (2011) vor, alle privaten Varianten zu einer Gesamt-Variante zusammenzufassen. Dies kann aber zu einer Verzerrung und so zu einem erhöhten Fehler 1. Art führen, wenn der Effekt der meisten kombinierten privaten Varianten neutral oder bi-direktional ist.

Die C-α-Methode wurde für die Untersuchung quantitativer Phänotypen und für die Einbeziehung von Kovariablen durch Wu et al. (2011a) weiterentwickelt (vgl. Abschnitt 2.3.14).

Algorithmus

1. Zur Stichprobe $\{1, \ldots, n\}$ seien die beiden Teilmengen der Kontrollen $N^u \subset \{1, \ldots, n\}$ und der Fälle $N^a \subset \{1, \ldots, n\}$ gegeben, so dass für die Anzahlen gilt $\#N^u = n^u$ und $\#N^a = n^a$ mit $n^a + n^u = n$. Schätze daraus den Anteil der Fälle innerhalb der Stichprobe mit $p_0 = n^a/n$.

2. Zähle für jede der Varianten $j \in \{1, \ldots, K\}$ der ROI die seltenen Allele in der Gesamtstichprobe $n_j = \sum_{i=1}^{n} x_{ij}$ sowie in der Gruppe der Fälle $n_j^a = \sum_{i \in N^a} x_{ij}$ unter Verwendung der Genotypkodierung gemäß Gleichung (2.1).

3. Betrachte die F verschiedenen Anzahlen aller seltenen Allele innerhalb der ROI für die gesamte Stichprobe, $z_f, f = 1, \ldots, F$.

4. Berechne die Teststatistik $T^{\text{C-}\alpha}$ gemäß

$$T^{\text{C-}\alpha} = \frac{\sum_{j=1}^{K}[(n_j^a - n_j p_0)^2 - n_j p_0(1-p_0)]}{\sum_{f=1}^{F}\sum_{j=1}^{K} I\{n_j = z_f\} \sum_{u=0}^{z_f}\left[(u - z_f p_0)^2 - z_f p_0(1-p_0)\right]^2 \text{Bin}_{z_f,p_0}(u)}$$

   Dabei ist $\text{Bin}_{z_f,p_0}(u)$ die Wahrscheinlichkeitsfunktion der Binomialverteilung mit den Parametern z_f und p_0 sowie I die Indikatorfunktion.

5. Schätze aus der Teststatistik $T^{\text{C-}\alpha}$ den rechtseitigen p-Wert aus der Standardnormalverteilung.

2.3.12 FPCA – Functional Principal Component Analysis

Luo et al. (2011) entwickelten die Gruppierungsmethode der funktionellen Hauptkomponentenanalyse (FPCA). Die FPCA-Methode basiert auf den Konzepten des kontinuierlichen Genommodells von Bickeböller und Thompson (1996) und nutzt die Methode der Hauptkomponentenanalyse (PCA). Die zugrundeliegende Idee der FPCA-Methode soll im Folgenden grob umrissen werden.

Das kontinuierliche Genommodell betrachtet die DNA bzw. die Genotyp-Informationen des gesamten Genoms als stetig und erlaubt daher die Projektion der Varianten $1, \ldots, K$ der ROI auf das stetige Intervall $[0, 1]$. Anschließend wird ein Integral s bestehend aus der kontinuierlichen Linearkombinantion bzgl. der normierten Gewichtsfunktion $\beta(t)$ und der Genotyp-Information $x(t)$ für jedes $t \in [0, 1]$ gebildet:

$$s = \int_{[0,1]} \beta(t)x(t)dt$$

Die Gewichtsfunktion $\beta(t)$ ist unbekannt und muss ermittelt werden. Dabei muss $\beta(t)$ die Bedingung erfüllen, dass die Varianz der Funktion s, $Var(s)$, maximal wird. Die Lösung dieser Gleichung führt auf ein Optimierungsproblem, das mit einer Hauptkomponentenanalyse (PCA) gelöst werden kann. Da die PCA für die Funktion $\beta(t)$ erfolgt, handelt es sich um eine funktionelle PCA. Mathematisch betrachtet entspricht das Verfahren einer Hauptachsentransformation bzw. einer Singulärwertzerlegung. Für eine detaillierte Darstellung der Methode sei auf (Fischer 1983; Luo et al. 2011) verwiesen.

Daraufhin wird das Optimierungsproblem in weitere Integralfunktionen, die sogenannten Hauptkomponentenfunktionen (Luo et al. 2011), überführt. Aus den Hauptkomponentenfunktionen werden Hauptkomponenten-Scores gebildet, die zur Berechnung der Teststatistik verwendet werden. Da die Hauptkomponentenfunktionen nicht in geschlossener Form lösbar sind, werden sie mit Hilfe eines Diskretisierungsansatzes gelöst, der im folgenden Algorithmus beschrieben wird.

Für den Fall, dass Varianten der betrachteten ROI im Kopplungsungleichgewicht sind, schlagen Luo et al. (2011) vor, die Genotyp-Informationen der Stichprobe bzgl. einer Fourier-Basis in entsprechende Koeffizienten zu zerlegen, die die gleiche Information wie die ursprünglichen Genotypdaten tragen.

Bei den in der FPCA-Methode verwendeten Gewichten ist nicht klar, wie diese möglicherweise protektive oder schädliche Varianten auf- oder abwerten, da die Gewichte aus den Hauptkomponenten entwickelt werden. Dadurch kann nicht genau differenziert werden, ob FPCA ein burden oder nicht-burden Test ist. Um die Teststärke zusätzlich zu erhöhen schlagen Luo et al. (2011) die Verwendung von Qualitätsparametern der Sequenzierung für jede Variante als Gewicht vor. Im Allgemeinen ist FPCA nicht auf die Untersuchung von ausschließlich seltenen Varianten beschränkt,

jedoch ist nur die Betrachtung eines dichotomen Phänotyps ohne Einbeziehung von Kovariablen möglich. Für die Zusammenfassung weiterer Eigenschaften wird auf Tabelle 2.1 verwiesen.

Für ROIs, für die die MAFs der Varianten innerhalb der ROI sehr stark schwanken, haben Luo et al. (2013) ihre Methode FPCA weiterentwickelt und einen Glättungsparameter für die ROI bzgl. der gegebenen Stichprobe in den Ansatz integriert. Auf diese Erweiterung wird im Folgenden aber nicht weiter eingegangen.

Algorithmus

1. Projiziere die Basenpaarpositionen $g_1, \ldots, g_K$ der Varianten einer ROI auf das Intervall $[0, 1]$ und ordne sie in aufsteigender Reihenfolge:

$$\bar{g}_k := \frac{g_k - g_1}{g_K - g_1}, \qquad k = 1, \ldots, K.$$

2. Definiere eine Fourier-Basis $\phi_d, d = 1, \ldots, D$, bestehend aus D Elementen. Je größer D ist, desto größer ist der Anteil der Varianz von s, die über das Modell erklärt werden kann. Werte die Fourier-Basis an den normierten Basenpaarpositionen $\bar{\boldsymbol{g}} = (\bar{g}_1, \ldots, \bar{g}_K)$ der ROI aus, so dass eine Fourier-transformierte ROI entsteht:

$$\boldsymbol{F} := \mathrm{FT}(\bar{\boldsymbol{g}})(\phi_{d,j})_{\substack{d=1,\ldots,D,\\ æ=1,\ldots,K}} \tag{2.32}$$

3. Es sei $\boldsymbol{X} = (x_{ij})_{\substack{i=1,\ldots,n,\\ æ=1,\ldots,K}}$ die Genotypmatrix der ROI in der Stichprobe. Berechne aus Fourier-Transformation $\boldsymbol{F}$ die zugehörige Koeffizientenmatrix $\boldsymbol{C} = (c_{di})_{\substack{d=1,\ldots,D,\\ i=1,\ldots,n}}$, so dass die Genotypmatrix $\boldsymbol{X}$ wie folgt geschätzt werden kann:

$$\hat{\boldsymbol{X}} = \boldsymbol{C}^\intercal \boldsymbol{F}$$

4. Zentriere jede Zeile der Koeffizientenmatrix $\boldsymbol{C}$ mit dem Mittelwert der jeweiligen Zeile, woraus eine zentrierte Koeffizientenmatrix $\overline{\boldsymbol{C}}$ entsteht.

$$\overline{\boldsymbol{C}} = (\bar{c}_{di})_{\substack{d=1,\ldots,D,\\ i=1,\ldots,n}} \quad \text{mit } \bar{c}_{di} = c_{di} - \tfrac{1}{n}\sum_{i=1}^{n} c_{di}$$

5. Führe eine Hauptkomponentenanalyse (PCA) bzgl. der zentrierten und transponierten Koeffizientenmatrix $\bar{\boldsymbol{C}}^{\intercal}$ durch, was auf eine Diagonalmatrix von Gewichten $\boldsymbol{W}^{\mathrm{FPCA}} = \operatorname{diag} \boldsymbol{w}$ mit $\boldsymbol{w} = (w_1, \ldots, w_D)$ führt, die als Ladungen bezeichnet werden.

6. Berechne die sogenannte Hauptkomponenten-Score-Matrix entsprechend

$$\boldsymbol{S} = \overline{\boldsymbol{C}}^{\intercal} \boldsymbol{W}^{\mathrm{FPCA}} \tag{2.33}$$

7. $\boldsymbol{S}$ besteht aus den Teilmatrizen $\boldsymbol{S}^a$ und $\boldsymbol{S}^u$, für die Fälle und die Kontrollen: $\boldsymbol{S^a} = (s_{id})_{i\in N^a, d=1,\ldots,D}$ bzw. $\boldsymbol{S^u} = (s_{id})_{i\in N^u, d=1,\ldots,D}$. Berechne durch Spaltenweise Mittelwertbildung den mittleren Hauptkomponenten-Score-Vektor $\bar{\boldsymbol{v}}^{\boldsymbol{a}} = (\bar{s}^a_1, \ldots, \bar{s}^a_D)^{\intercal}$ für die Fälle und $\bar{\boldsymbol{s}}^{\boldsymbol{u}} = (\bar{s}^u_1, \ldots, \bar{s}^u_D)^{\intercal}$ für die Kontrollen.

8. Die Abstände der mittleren Hauptkomponenten-Scores aus Fällen und Kontrollen werden solange aufsummiert bis die maximale Varianz erreicht ist. Berechne die Teststatistik

$$T^{\mathrm{FPCA}} = \left(\frac{1}{n^a} + \frac{1}{n^u}\right)^{-1} \sum_{d=1}^{D} \frac{1}{v_d} (\bar{s}_d - \bar{s}_d)^2$$

mit

$$v_d = \frac{1}{n^a + n^u - 2} \left(\sum_{i\in N^a} (s^a_{di} - \bar{s}^a_d)^2 + \sum_{i\in N^u} (s^u_{di} - \bar{s}^u_d)^2 \right)$$

9. Da T^{FPCA} asymptotisch zentral-χ^2_D-verteilt ist, mit D Freiheitsgraden, schätze den einseitigen p-Wert entsprechend aus einer χ^2_D-Verteilung mit D Freiheitsgraden.

2.3.13 PWST – P-value Weighted Sum Test

Der p-Wert-gewichtete Summentest (PWST) wurde von Zhang et al. (2011) entwickelt und gehört zu den ersten Gruppierungsmethoden, die bi-direktionale Effekte innerhalb einer ROI in Betracht ziehen. Dieser Ansatz ist ähnlich der ASUM Methode (Abschnitt 2.3.9). Der Unterschied zwischen den Methoden ist, dass bei ASUM zunächst die Effektstärken einzeln für jede Variante der ROI über eine multiple Regression geschätzt werden und anschließend die Allele für die durch das Verfahren ein protektiver Effekt geschätzt wurde, in ein häufiges Allel umkodiert werden. Mit PWST ist sowohl die Untersuchung eines quantitativen als auch eines qualitativen Phänotyps möglich. Der ursprüngliche Ansatz sah keine die Berücksichtigung von Kovariablen vor (Zhang et al. 2011). Diese können aber leicht in das Regressionsmodell (wie bei den Gleichungen (2.3) und (2.4) beschrieben) eingefügt werden. Weitere Eigenschaften von PWST sind in Tabelle 2.1 zusammengefasst.

Aufgrund der enorm hohen Rechenzeit des PWST wird in dieser Arbeit nur das Szenario eines quantitativen Phänotyps ohne Berücksichtigung von Kovariablen untersucht.

Algorithmus

1. Betrachtet wird ein vereinfachtes lineares Regressionsmodell ohne Kovariablen mit Genotyp-kodierten Varianten $j = 1, \ldots, K$ für ein Individuum $i \in \{1, \ldots, n\}$

$$y_i = \tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_i + \varepsilon_i$$

   mit der Phänotypausprägung y_i, dem Achsenabschnitt $\tilde{\beta}_0$, dem Genotyp-Vektor $\boldsymbol{x}_i = (x_{i1}^G, \ldots, x_{iK}^G)^\intercal$, vgl. Gleichung (2.2), und dem Fehlerterm ε_i (vgl. das allgemeine Regressionsmodell in Gleichung (2.3)). Schätze die Einzeleffekte $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_K)^\intercal$ und den zugehörigen Standardfehler $S.E._{\tilde{\beta}_j}$ für die Varianten $j = 1, \ldots, K$.

2. Berechne für jede Variante $j \in \{1, \ldots, K\}$ die Teststatistik aus dem Effekt und dem zugehörigen Standardfehler aus Punkt **1.**:

$$T_{0,j}^{\text{PWST}} = \frac{\tilde{\beta}_j}{S.E._{\tilde{\beta}_j}}, \quad j = 1, \ldots, K.$$

Die zugehörigen linksseitigen p-Werte $p_{0,j}$ schätzt man mit Hilfe eines Wald-Tests (Bera und Bilias 2001; Engle 1984).

3. Berechne aus den in Punkt **2.** geschätzten p-Werten Effektrichtungs-spezifische Gewichte:

$$w_j = 2[p_{0,j} - 0{,}5], \quad j = 1, \ldots, K. \tag{2.34}$$

4. Betrachtet wird nun ein zweites Regressionsmodell mit den zuvor berechneten Gewichten $\boldsymbol{w} = (w_1, \ldots, w_K)^\intercal$ und einem gemeinsamen Effekt β für alle Varianten für jedes Individuum $i \in \{1, \ldots, n\}$:

$$y_i = \beta_0 + \beta \boldsymbol{w}^\intercal \boldsymbol{x}_i + \varepsilon_i$$

5. Teste die Nullhypothese $H_0 : \beta = 0$ mittels eines Likelihood-Quotiententests (Bera und Bilias 2001), welcher den p-Wert p_0 liefert.

6. Permutiere den quantitativen Phänotypvektor L mal und führe die Punkte **4.** und **5.** mit den permutierten Phänotypvektoren $\boldsymbol{y_\ell}, \ell = 1, \ldots, L$, und den Gewichten $\boldsymbol{w}$ aus dem nicht-permutierten Datensatz aus Punkt **3.** erneut durch.

7. Bestimme den p-Wert p^{PWST} über den Anteil an p-Werten bzgl. der permutierten Daten, welche mindestens so groß sind wie der Original p-Wert p_0:

$$p^{\text{PWST}} = \frac{1}{L+1}\Big(\sum_{\ell=1}^{L} I\{p_\ell \geq p_0\} + 1\Big) \tag{2.35}$$

Dabei sind $p_\ell, \ell = 1, \ldots, L$, die p-Werte bzgl. der permutierten Daten.

2.3.14 SKAT – Sequencing Kernel Association Test

Der Kern-basierte Sequenzierungs-Assoziationstest (SKAT) ist ein Varianz-Komponententest und wurde von Wu et al. (2011a) entwickelt. Er stellt eine Verallgemeinerung des C-α-Ansatzes von Neale et al. (2011) (vgl. Abschnitt 2.3.11), des *Sum of Squared Score (SSS)* Tests von Pan (2009) und dem *Haplotype Association Test* von Tzeng und Zhang (2007) dar. SKAT basiert auf einem Regressionsansatz wie in Abschnitt 2.2.4 beschrieben und ist sowohl auf dichotome als auch auf quantitative Phänotypen unter Betrachtung von Kovariablen anwendbar. Im Wesentlichen beruht SKAT auf der Annahme eines linearen Zusammenhangs zwischen einem Phänotyp und einer ROI mit voneinander verschiedenen Einzel-Effekten β_j der darin enthaltenen Varianten $j \in \{1, \ldots, K\}$ (Burkett und Greenwood 2013). Der Unterschied zu anderen den Gruppierungsmethoden, die diesen Ansatz verfolgen, ist die Annahme, dass die Einzel-Effekte β_j der Varianten einer Verteilung mit Mittelwert 0 und Varianz $w_j\gamma$, $j \in \{1, \ldots, K\}$ folgen. Dabei ist w_j ein definiertes Varianten-spezifisches Gewicht und γ ist eine Varianzkomponente, die unter der Nullhypothese, dass keine Assoziation vorliegt, den Wert 0 annimmt. Im Falle eines dichotomen Phänotyps und Gewichten $w_j = 1$, $j \in \{1, \ldots, K\}$ sind die Teststatistiken von SKAT und C-α identisch (Abschnitt 2.3.11).

SKAT ist robuster als andere Gruppierungsmethoden gegenüber ROIs mit Varianten unterschiedlicher Effekt-Richtungen. Dies wird deutlich am Aufbau der Teststatistik, in der Ähnlichkeiten in der Genotypstruktur für alle Paare von Individuen in der betrachteten Stichprobe gemessen werden. Diese werden je Variante gewichtet und für die ROI zu einem Score aufsummiert. Skat ist nach der Klassifizierung von Derkach et al. (2014) (Abschnitt 2.2.5.2) entsprechend der Berechnung der Teststatistik T^{SKAT} in die Klasse quadratischer Teststatistiken einzuordnen. Weitere Eigenschaften von SKAT sind in Tabelle 2.1 zusammengefasst.

In dieser Arbeit wird SKAT sowohl für die Anwendung auf quantitative als auch dichotome Phänotypen jeweils mit und ohne Kovariablen untersucht. Mit dem folgenden Algorithmus wird nur der Fall eines quantitativen Phänotyps unter Verwendung von Kovariablen vorgestellt. Der Fall eines dichotomen Phänotyps lässt sich leicht verallgemeinern, vgl. Abschnitt 2.2.4.

Im Algorithmus werden u.a. Eigenwerte auf Basis von positiv-semidefinten Matrizen berechnet und weitere Methoden der numerischen linearen Algebra verwendet. Es wird im Folgenden angenommen, dass diese Voraussetzungen gemäß der Originalarbeit von Wu et al. (2011a) i.d.R. erfüllt sind. Für mathematische Details und Hintergründe wird daher auf die Originalarbeit der Autoren verwiesen.

Algorithmus

1. Betrachtet wird für ein Individuum $i \in \{1, \ldots, N\}$ das lineare Regressionsmodell:

$$y_i = \beta_0 + \boldsymbol{\beta}^\intercal \boldsymbol{x}_i + \boldsymbol{\alpha}^\intercal \boldsymbol{c}_i + \varepsilon_i \tag{2.36}$$

mit dem Achsenabschnitt β_0, voneinander unabhängigen Effekten $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^\intercal$, wobei jeder Effekt β_j einer Verteilung mit Mittelwert 0 und Varianz $w_j\gamma$ folgt, $\boldsymbol{x}_i$ ist der Genotypvektor, $\boldsymbol{c_i}$ bezeichnet den Vektor der Kovariablen mit Koeffizienten $\boldsymbol{\alpha}$ und ε_i ist der Fehlerterm. Die entsprechende Nullhypothese $H_0 : \beta = \beta_1 = \ldots = \beta_K = 0$ ist äquivalent zu der Annahme, dass $\gamma = 0$ ist.

2. Wähle ein Model für die Gewichte $\boldsymbol{w}$. Wu et al. (2011a) schlagen dafür als Ausgangsgewichte die Dichte der Betaverteilung (Abramowitz und Stegun 1965) vor $\sqrt{w_i} = \text{Beta}(\text{MAF}_i; a_1, a_2)$ mit $a_1 = 1$ und $a_2 = 25$. Diese Wahl führt dazu, dass den Varianten mit MAF $\leq$ 0,01 ein größeres und denen mit einer höheren MAF ein kleineres Gewicht zukommt. Entsprechend einer abweichenden Annahme über die untersuchte Assoziation können andere Werte für a_1 und a_2 sinnvoll sein. Zum Beispiel liefert die Wahl $a_1 = a_2 = 0{,}5$ die Gewichte $w_j = 1/(\text{MAF}_j(1 - \text{MAF}_j))$, was den Gewichten im Ansatz WSS von Madsen und Browning (2009) sehr ähnelt (Abschnitt 2.3.6).

3. Die Kernfunktion der Teststatistik kann je nach Annahme der zugrundeliegenden Effekte verschieden zusammengesetzt sein. Wu et al. (2011a) schlagen die Verwendung einer der folgenden Kernfunktionen mit den zuvor gewählten Gewichten $\boldsymbol{w}$ vor:

- Gewichteter linearer Kern

$$\boldsymbol{K}(x_i,x_{i'}) = \sum_{j=1}^{K} w_j x_{ij} x_{i'j}$$

  Dieser ist der einfachste und intuitivste Kern, der auf der Annahme linearer genetischer Effekte basiert.

- Gewichteter quadratischer Kern

$$\boldsymbol{K}(x_i,x_{i'}) = \Big(1 + \sum_{j=1}^{K} w_j x_{ij} x_{i'j}\Big)^2$$

  Dieses Modell impliziert, dass nur Haupteffekte vorliegen.

- Gewichteter IBS-Kern

$$\boldsymbol{K}(x_i,x_{i'}) = \sum_{j=1}^{K} w_j \mathrm{IBS}(x_{ij}, x_{i'j})$$

  Dabei gibt IBS(x_{ij},$x_{i'j}$) (Identical by state) die Anzahl der identischen Allele je Individuenpaar (i, i') für eine Variante $j \in \{1, \ldots, K\}$ an.

4. Berechne mit Hilfe der gewählten Kernfunktion die Matrix $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\intercal$. $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ ist der gewichtete, lineare Kern, und gibt in jedem Element die genetische Ähnlichkeit für jedes Individuenpaar (i, i'), $i, i' \in \{1, \ldots, n\}$, der Stichprobe an. $\boldsymbol{X}$ ist eine $n \times K$-Matrix, die an Position (i, j) für jedes Individuum i die Anzahl der seltenen Allele der Variante j enthält und $\boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_K)$ eine Diagonal-Matrix, die die einzelnen Gewichte der K Varianten auf der Hauptdiagonalen enthält.

5. Schätze unter H_0 für alle Individuen $i \in \{1, \ldots, n\}$ die erwartete Phänotypausprägungen $\hat{\mu}_i = \hat{\beta}_0 + \hat{\boldsymbol{\alpha}}^\intercal \boldsymbol{c_i}$ der Stichprobe, sowie den Achsenabschnitt $\hat{\beta}_0$, die Effekte $\hat{\boldsymbol{\alpha}}$ der Kovariablen und die Varianz $\hat{\sigma}^2$ des residualen Fehlerterms.

6. Berechne die Teststatistik als Varianz-Komponenten-Score:

$$T^{\mathrm{SKAT}} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\intercal \boldsymbol{K} (\boldsymbol{y} - \hat{\boldsymbol{\mu}})$$

7. Berechne die Eigenwerte $\lambda_1, \ldots, \lambda_n$ von

$$\boldsymbol{P}_0^{1/2} \boldsymbol{K} \boldsymbol{P}_0^{1/2}$$

mit der positiv-definiten Matrix

$$\boldsymbol{P_0} = \boldsymbol{V^{-1}} - \boldsymbol{V^{-1}} \tilde{\boldsymbol{X}} (\tilde{\boldsymbol{X}}^\intercal \boldsymbol{V^{-1}} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^\intercal \boldsymbol{V^{-1}}, \tag{2.37}$$

und $\tilde{\boldsymbol{X}} := [\mathbf{1}, \boldsymbol{X}] \in \mathbb{R}^{n \times (K+1)}$. Dabei ist $\mathbf{1}$ ein Vektor der Länge n mit dem Wert 1 in jedem Element, $\boldsymbol{V^{-1}} := 1/\hat{\sigma^2} \boldsymbol{I}$ enthält die geschätzte Varianz $\hat{\sigma^2}$ aus **5.** und $\boldsymbol{I}$ ist die Identitätsmatrix der Größe $n \times n$. Für mathematische Details wird auf die Originalarbeit von Wu et al. (2011a) verwiesen.

8. Schätze den p-Wert zu T^{SKAT} aus einer gemischten χ^2-Verteilung der Form $\sum_{\ell=1}^{\Lambda} \lambda_\ell \chi^2_{1,\ell}$, wobei $\lambda_1, \ldots, \lambda_\Lambda$ die Eigenwerte ungleich 0 sind, mit $\Lambda \leq K$.

2.3.15 SKAT-O - Optimal Unified Sequencing Kernel Association Test

Die optimierte und vereinheitlichte SKAT-Methode (SKAT-O) wurde von Lee, Emond et al. (2012) als eine direkte Weiterentwicklung der SKAT-Methode (Abschnitt 2.3.14) vorgeschlagen. Das Ziel war es, sowohl die Beschränkungen der Burden- als auch der Nicht-burden-Methoden zu überwinden, weiterhin aber die jeweiligen Stärken der Ansätze zu nutzen. Die grundlegende Idee von SKAT-O ist es, eine optimale Kombination aus einem Burden- und einem Nicht-burden-Test zu finden, so dass die optimale Teststatistik erzeugt wird, in dem Sinne, dass der p-Wert aus den kombinierten Methoden minimal wird. Lee, Emond et al. (2012) schlagen eine Linearkombination aus der SKAT Teststatistik (Abschnitt 2.3.14) und einer beliebigen Nicht-Burden-Teststatistik vor. Da SKAT-O stark von der SKAT-Methode abhängt, basiert auch diese Methode auf einem Regressionsansatz wie in Abschnitt 2.2.4 eingeführt. SKAT-O kann sowohl für kontinuierliche und als auch dichotome Phänotypen und unter Einbeziehung von Kovariablen verwendet werden (vgl. Tabelle 2.1). Durch die Verwandtschaft zu SKAT ist auch SKAT-O robust gegenüber bi-direktionalen Effekten innerhalb einer ROI. Da in vielen aktuellen Studien die Stichprobenumfänge

noch relativ klein sind, insbesondere im Zusammenhang mit seltenen Varianten, und somit die Teststärke oft gering ist, haben Lee, Emond et al. (2012) ihren Ansatz um einen analytischen Ansatz erweitert, um die Verteilung unter der Nullhypothese besser schätzen zu können. Hierfür wird zusätzlich sowohl die Varianz als auch die Kurtosis unter der Nullhypothese geschätzt um die asymptotische Verteilung unter H_0 zu approximieren.

In der folgenden Beschreibung des Algorithmus wird lediglich der Fall eines kontinuierlichen Phänotyps unter Verwendung von Kovariablen betrachtet. Auf die Beschreibung des analytischen Ansatzes wird hier verzichtet, dieser wird in Lee, Emond et al. (2012) beschrieben.

Algorithmus

1. Betrachtet wird für ein Individuum $i \in \{1, \ldots, n\}$ das gleiche lineare Regressionsmodell wie in Gleichung (2.36):

$$y_i = \beta_0 + \boldsymbol{\beta}^\intercal \boldsymbol{x}_i + \boldsymbol{\alpha}^\intercal \boldsymbol{c}_i + \varepsilon_i$$

   mit der Ausprägung y_i des quantitativen Phänotyps, dem Achsenabschnitt β_0, voneinander unabhängigen zufälligen Effekten $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^\intercal$, wobei jeder Effekt β_j, $j = 1, \ldots, K$, einer Verteilung mit Mittelwert 0 und Varianz $w_j^{\mathrm{SKAT}}\gamma$ folgt. Dabei sind $\boldsymbol{w}^{\mathrm{SKAT}} = (w_1^{\mathrm{SKAT}}, \ldots, w_K^{\mathrm{SKAT}})^\intercal$ variantenbasierte Gewichte und die Genotypen $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK})^\intercal$ der betrachteten ROI, $\boldsymbol{c}_i$ ist der Vektor der Kovariablen und $\boldsymbol{\alpha}$ die entsprechenden Koeffizienten und ε_i der Fehlerterm.

2. Berechne die zugehörige Teststatistik T^{SKAT} (vgl. Abschnitt 2.3.14).

3. Betrachte zusätzlich das folgende lineare Regressionsmodell für Individuum $i \in \{1, \ldots, n\}$:

$$y_i = \tilde{\beta}_0 + \tilde{\beta}\boldsymbol{x}_i^\intercal \boldsymbol{w}^{\mathrm{Burden}} + \tilde{\boldsymbol{\alpha}}^\intercal \tilde{\boldsymbol{c}}_i + \tilde{\varepsilon}_i$$

   mit der Phänotypausprägung y_i, dem Achsenabschnitt $\tilde{\beta}_0$, dem kollektiven Effekt $\tilde{\beta}$ der ROI, dem Genotyp-Vektor $\boldsymbol{x}_i = (x_{i1}^G, \ldots, x_{iK}^G)^\intercal$, dem Vektor

von Gewichten $\boldsymbol{w}^{\text{Burden}} = (w_1^{\text{Burden}}, \ldots, w_K^{\text{Burden}})^\intercal$ für jede der Varianten $j = 1, \ldots, K$ der ROI und dem zufälligen Fehler $\tilde{\varepsilon}_i$.

4. Berechne die allgemeine Burden-Score-Teststatistik (Derkach et al. 2014; Lee, Emond et al. 2012):

$$T^{\text{Burden}} = \left(\sum_{i=1}^{n} (y_i - \hat{y}_i) \Big(\sum_{j=1}^{K} w_j^{\text{Burden}} x_{ij} \Big) \right)^2. \tag{2.38}$$

Dabei ist $\hat{y}_i$ die aus dem Modell in Punkt **3.** geschätzte Phänotypausprägung von Person i.

5. Berechne für ein zuvor festgelegtes, möglichst feines Gitter $0 \leq \rho_1 \leq \cdots \leq \rho_R \leq 1$ die gewichteten Mittel aus der Burden- und der SKAT-Teststatistik:

$$T_{\rho_r}^{\text{SKAT-O}} = \rho_r T^{\text{Burden}} + (1 - \rho_r) T^{\text{SKAT}}.$$

6. Berechne die Teststatistik aus dem Maximum der zu den gewichteten Mitteln aus Punkt **5.** gehörenden Teststatistiken:

$$T^{\text{SKAT-O}} = \max_{r \in \{1, \ldots, R\}} T_{\rho_r}^{\text{SKAT-O}}. \tag{2.39}$$

7. Berechne den p-Wert zu $T^{\text{SKAT-O}}$. Lee, Wu et al. zeigen in verschiedenen Arbeiten (Lee, Wu et al. 2012; Lee, Emond et al. 2012), dass $T^{\text{SKAT-O}}$ asymptotisch einer Kombination aus einer adjustierten χ^2-Mischverteilung und einer einzelnen χ^2-Verteilung mit einem Freiheitsgrad folgt. Zusätzlich konnten die Autoren zeigen, dass der p-Wert analytisch mit Hilfe einer eindimensionalen Integration geschätzt werden kann.

3 Simulationsstudie

Durch eine Anwendung der Methoden auf Simulationsdaten soll ein neutraler Vergleich aller in dieser Arbeit betrachteten Gruppierungsansätze möglich werden. Der zugrundeliegende Datensatz wurde im Rahmen des *Genetic Analysis Workshop 17* (GAW17) simuliert, worauf in Abschnitt 3.1 im Detail eingegangen wird (Almasy et al. 2011). Um die Leistung der Gruppierungsmethoden bei Verwendung verschiedener Filterkriterien untersuchen zu können, werden in Abschnitt 3.2 ROIs gebildet. Für diese wird die Assoziation sowohl bzgl. eines binären als auch eines quantitativen Phänotyps unter Berücksichtigung verschiedener Kovariablen, für alle entsprechenden Gruppierungsmethoden untersucht, wie in 3.3 beschrieben.

In vielen der vorgestellten Gruppierungsmethoden wird der p-Wert empirisch mittels Permutationsansatz geschätzt. Da dies sehr rechenintensiv ist, wird ein vierstufiges Schätzverfahren zur Ermittlung des empirischen p-Werts verwendet, das in Abschnitt 3.4 näher beschrieben wird. Die Vorgehensweise zur Bestimmung des Fehlers 1. Art und der Teststärke wird in Abschnitt 3.5 beschrieben. Die Simulationsergebnisse bzgl. der verschiedenen ROIs, Phänotypen und Kovariablen werden anschließend in Abschnitt 3.6 dargelegt. Des Weiteren wird dort die Teststärke für die einzelnen assoziierten ROIs über die insgesamt 200 Replikate aus den verschiedenen Untersuchungsszenarien bzgl. des Anteils der verursachenden Varianten, der kumulierten MAF und des kumulierten genetischen Effekts der Varianten einer ROI betrachtet. Schließlich erfolgt in Abschnitt 3.7 die Interpretation der Ergebnisse.

3.1 Simulationsaufbau

Im Rahmen des *Genetic Analysis Workshops 17* (GAW17)(Almasy et al. 2011) wurden Daten eines *Mini-Exoms* simuliert, um neu entwickelte statistische Methoden daran testen zu können. Die Mini-Exom-Daten basieren auf den Real-Daten der *Pilot-Studie 3 des 1000 Genomes Projects* (1000 Genomes Project Consortium et al. 2010). Mit Hilfe dieser Daten wurde eine häufige komplexe Krankheit mit zusätzlichen Risikofaktoren simuliert.

Insgesamt sind in diesem Datensatz 24,487 autosomale SNPs bzw. Varianten auf Basis eines männliches Referenzgenoms (RefSeq36) des *National Center for Biotechnology Information* (Pruitt et al. 2007) 3,205 Genen zugeordnet. Innerhalb eines Gens befinden sich zwischen 1 und 231 Varianten bzw. SNPs, von denen 74% eine MAF $\leq$ 0,01 und nur 12,8% eine MAF $\geq$ 0,05 haben.

Es wurden 697 unverwandte Personen simuliert, die ursprünglich aus dem 1000 Genomes Project stammen und verschiedenen Populationen angehören. Davon sind 327 männlich und 370 weiblich. Das Alter der Personen liegt zwischen 16 und 91 Jahren.

Der verwendete Phänotyp ist eine häufige Krankheit mit einer Prävalenz von 30%. Neben dem Fall-Kontroll-Status wurden auch drei normalverteilte Risikofaktoren $Q1$, $Q2$, $Q4$ und ein Raucherstatus mit einer Prävalenz von 25% simuliert. Es wurden 200 Replikate des Phänotyps gebildet, bei denen die Genotyp-, Alters- und Geschlechtsdaten festgehalten wurden. Zur Bildung einer Gengruppe, die für einen biologischen Prozess verantwortlich ist, wurden zur Simulation des komplexen Phänotyps im wesentlichen SNPs bzw. Varianten aus dem vaskulären endothelialen Wachstumsfaktor (engl. Vascular Endothelial Growth Factor, VEGF) Pathway verwendet.

Der Risikofaktor $Q1$ wird durch neun Gene mit insgesamt 39 SNPs des VEGF Pathways beeinflusst. Er ist unter Rauchern stets höher und auch mit dem Alter positiv korreliert.

Für den Risikofaktor $Q2$ werden keine Gene eines bestimmten Pathways gewählt, sondern im Wesentlichen 13 Gene mit insgesamt 72 SNPs, die das kardiovaskuläre

Krankheitsrisiko oder das Risiko für eine Entzündung des Herzmuskels beeinflussen. Das Vorkommen seltener Allele ist mit einem höheren Wert von $Q2$ verbunden. $Q2$ ist mit $Q1$ und der latenten Krankheitsanfälligkeit, nicht aber mit Alter, Geschlecht oder Raucherstatus korreliert.

Der Risikofaktor $Q4$ hat eine protektive Wirkung und wird durch keine der vorhandenen SNPs beeinflusst. $Q4$ ist geringer unter Rauchern und Frauen und sinkt mit steigendem Alter.

Die 15 Gene, die eine latente Krankheitsanfälligkeit bestimmen, umfassen 51 SNPs des VEGF-Pathways. Sie unterscheiden sich von denen, die $Q1$ beeinflussen. Die mittlere Anfälligkeit ist stets mit dem seltenen Allel assoziiert, ist bei Rauchern höher und steigt mit dem Alter.

Alle assoziierten Varianten und SNPs sind nicht-synonym in der Kodierung der zugehörigen Aminosäure. Die Effektstärken der einzelnen Varianten bzw. SNPs wurden abgeleitet aus den durch *SIFT* (Ng und Henikoff 2003) und *PolyPhen* (Gorlov et al. 2008) angegebenen Wahrscheinlichkeiten, dass die entsprechende Variante schädlich ist. Die meisten Varianten bzw. SNPs hatten Effektstärken < 1 (Almasy et al. 2011). In der Simulation des Fall-Kontroll-Status wurde die Populationszugehörigkeit nicht als Einflussfaktor berücksichtigt. Mit der Ausnahme von wenigen privaten Mutationen wiesen nur sehr wenige SNPs ein Kopplungsungleichgewicht auf (Almasy et al. 2011).

Um den assoziierten Phänotyp zu simulieren wurde mit Hilfe aller Faktoren, Q_1, Q_2, Q_4 und der latenten Krankheitsanfälligkeit, ein multifaktorielles *Anfälligkeits-Schwellenwertmodell* (engl. liability threshold model) entwickelt, so dass sich die Krankheitsanfälligkeit mit jedem vorhandenen Risikofaktor und jeder Kopie des seltenen Allels additiv erhöht. Der Fall-Kontroll-Status $AFFECTED$ wird darüber definiert, dass bzgl. des Phänotyps die oberen 30% als Fälle und die übrigen 70% als Kontrollen deklariert wurden. Die Variabilität des Phänotyps in den verschiedenen Replikaten wird nur über die Residuen der Genotyp- und der Umweltkomponenten simuliert. Die Residuen der Genotypkomponenten sind stark mit $Q1$, $Q2$ und der latenten Krankheitsanfälligkeit korreliert, die Residuen der Umweltkomponenten hingegen nur schwach.

3.2 Filterszenarien – Die Regionen von Interesse

Zur Untersuchung der Leistungsfähigkeit der hier betrachteten Gruppierungsmethoden bzgl. der gegebenen Genotypdaten bestehend aus 3,205 Genen in 24,487 autosomalen Varianten bzw. SNPs wurden verschiedene Regionen von Interesse gebildet. Dabei wurden die bereits durch das GAW17-Team annotierten Informationen und Filterkriterien der Frequenz des seltenen Allels und der Funktionalität der enthaltenen Varianten verwendet.

Basis für die Bildung der ROIs zur Gruppierung der Varianten waren die funktionellen Einheiten der Gene. Ausgehend von diesen Einheiten wurden zur Filterung von seltenen Varianten entweder die MAF-Grenze 0,01 oder die MAF-Grenze 0,05 angelegt. Nach dem Filterschritt lagen bereits die ROIs für der ersten Szenarien vor: Alle Varianten eines Gens mit einer MAF von höchstens 0,01 bzw. 0,05. Zur Konstruktion zweier weiterer Untersuchungsszenarien wurden nur die Varianten für die ROI berücksichtigt, die die zugehörige Aminosäure nicht-synonym kodieren. Damit der Ansatz der Gruppierung von Varianten erfüllt war, gilt für alle betrachteten Szenarien, dass nur die ROIs in die Analyse aufgenommen wurden, die *mindestens zwei* Varianten enthalten. Insgesamt ergeben sich schließlich vier Filter bzw. Untersuchungsszenarien, die in Tabelle 3.1 zusammengefasst sind.

Tabelle 3.1: Zusammenfassung der untersuchten Szenarien der GAW17-Simulationsdaten; Filterung auf Basis der funktionellen Einheit des Gens bzgl. der Frequenz des seltenen Allels (MAF) und der Funktionalität der enthaltenen Varianten; angegeben sind zusätzlich die Anzahl (#) der Regionen von Interesse sowie die darin enthaltenen Varianten.

Szenario	MAF	Varianten	# ROI	# Varianten
1	0,05	Gen-basiert	1328	13747
2	0,01	Gen-basiert	1140	11677
3	0,05	Nicht-synonym	1014	8182
4	0,01	Nicht-synonym	909	7085

3.3 Phänotypen und Kovariablen

Die meisten hier betrachteten Gruppierungsmethoden sind für die Untersuchung der Assoziation einer ROI mit einem dichotomen Phänotyp (i.d.R. Fall-Kontroll-Status) entwickelt worden. In einer ersten Untersuchung werden daher alle Methoden mit Ausnahme des VT-Ansatzes von Price et al. (2010) und des PWST-Ansatzes von Zhang et al. (2011) hinsichtlich des Fall-Kontroll-Status in allen Szenarien der ROIs verglichen. Für VT und PWST wird aufgrund der langen Rechenzeit nur der quantitative Phänotyp untersucht.

In einer zweiten Untersuchung werden alle Methoden, die auch die Betrachtung eines quantitativen Phänotyps zulassen (vgl. Tabelle 2.1) und für die ein entsprechend implementiertes Software-Programm verfügbar war, für den Risikofaktor $Q2$ (Abschnitt 3.1), der hier als Phänotyp verwendet wird, miteinander verglichen.

Einige wenige Methoden erlauben die Einbeziehung von Kovariablen und liegen in einem geeigneten Software-Programm vor. Daher werden in dieser Arbeit in einer dritten und vierten Untersuchung die binären Kovariablen *Geschlecht* und *Raucherstatus* sowie die quantitativen Variablen $Q1$ und $Q4$ sowohl für den Fall-Kontroll-Status als auch für den quantitativen Phänotyp $Q2$ in die Untersuchung einbezogen.

3.4 Schätzung des p-Werts durch Permutation

Ist T_0 die Teststatistik eines statistischen Tests bzgl. einer gegebenen Stichprobe, dann erhält man den unverzerrten, empirischen p-Wert p zu einer festen Anzahl von Permutationen N über den Anteil der Teststatistiken auf Basis von permutierten Daten, der mindestens so groß ist wie T_0:

$$p = \frac{1}{N+1}\Big(\sum_{i=1}^{N} I\{T_i \geq T_0\} + 1\Big). \tag{3.1}$$

Dabei ist I die Indikatorfunktion, die den Wert 1 annimmt falls die Bedingung in den Klammern erfüllt ist und den Wert 0, falls nicht.

Die Schätzung des p-Werts durch Permutation ist sehr rechenintensiv, insbesondere wenn der zugrundeliegende wahre p-Wert sehr klein ist und mit einer hohen Konfidenz geschätzt werden soll. Des Weiteren erhöht sich die Rechenzeit drastisch, wenn beispielsweise eine genomweite Analyse durchgeführt wird, d.h. alle ca. 30.000 Gene des Genoms betrachtet werden.

Sieben der in dieser Arbeit untersuchten Gruppierungsmethoden schätzen den p-Wert mittels Permutation (vgl. Tabelle 2.1). Aufgrund der relativ hohen Anzahl an ROIs in den einzelnen Szenarien (vgl. Tabelle 3.1) ist die Rechenintensität für die Permutationsansätze sehr hoch. Für die Leistungsermittlung der Gruppierungsmethoden sind insbesondere die „kleinen“ , signifikanten p-Werte von Bedeutung und sollen möglichst genau geschätzt werden. Um die Rechenzeit dennoch deutlich zu verkürzen, wird ein 4-stufiges Schätzverfahren angewendet. Hierbei wird die Anzahl der Permutationen sukzessive von 2000 auf 10.000, 100.000 und schließlich 400.000 erhöht. Für jede der Stufen wird das 95%-ige Konfidenzintervall nach Wilson (1927) auf Basis der Anzahl der Permutationen der jeweiligen Stufe für ein α-Niveau von 5% geschätzt. In den einzelnen Stufen werden die p-Werte für jede ROI des betrachteten Szenarios gemäß Gleichung (3.1) empirisch geschätzt. Für die ROIs, für die der geschätzte p-Wert unterhalb der oberen Grenze des Konfidenzintervalls der jeweiligen Stufe liegt, werden zusätzliche Permutationen durchgeführt, um die p-Werte in der nächsthöheren Permutationsstufe genauer zu schätzen.

Die maximale Anzahl von Permutationen in der 4. Stufe ergibt sich wie folgt: Nach Bradley (1978) gilt ein statistischer Test als robust, wenn sich der geschätzte p-Wert im Intervall $[\frac{1}{2} \cdot \alpha, \frac{3}{2} \cdot \alpha]$ befindet. Um möglichst belastbare p-Werte in der letzten Permutationsstufe zu schätzen, wird ausgehend von einem nominellen α-Niveau von 0,05 ein *Bonferroni-adjustiertes* α^{B} berechnet mit $\alpha^{\mathrm{B}} = 0{,}05/1328 = 3{,}77 \cdot 10^{-5}$ berechnet. Es ergibt sich aus der maximalen Anzahl von betrachteten ROIs in Szenario 4 (1328, vgl. Tabelle 3.1). Anschließend wird für die rechte Intervallgrenze $\frac{3}{2} \cdot \alpha^{\mathrm{B}}$ abhängig von der Anzahl der Permutationen das 95%-ige Konfidenzintervall nach Wilson (1927) berechnet. Die maximale Anzahl an Permutationen für die Simulationsstudie wird so gewählt, dass das ursprüngliche α^{B} noch immer vom

Konfidenzintervall nach Wilson überdeckt wird. In unserem Fall beträgt sie 400.000 und wird für alle Filterszenarien von gebildeten Regionen von Interesse gewählt.

3.5 Schätzung des Fehlers 1. Art und der Teststärke

Zur Beurteilung der Leistung bzw. Validität der hier betrachteten Gruppierungsmethoden werden der Fehler 1. Art und die Teststärke bzgl. der Simulationsdaten aus dem GAW17 über 200 Replikate empirisch geschätzt. Für die Bewertung der Teststärke werden die minimale Teststärke und die empirische Teststärke als Kennzahlen untersucht und werden zusammen mit dem Fehler 1. Art für jedes der in Tabelle 3.1 beschriebenen Szenarien geschätzt.

3.5.1 Der empirische Fehler 1. Art

Für einen Hypothesentest liegt ein Fehler 1. Art vor, falls eine Nullhypothese H_0 abgelehnt wird, obwohl sie wahr ist. Im Falle einer Simulationsstudie heißt das, dass eine nicht-assoziierte ROI auf Basis eines signifikanten p-Werts als krankheitsverursachend angenommen wird.

In dieser Arbeit wird für jede Gruppierungsmethode ein über die 200 Replikate gemittelter empirischer Fehler 1. Art geschätzt. Dabei wird der Fehler 1. Art für ein Replikat über den Anteil der nicht-assoziierten ROIs ermittelt, für die zu einem zuvor festgelegten Signifikanzniveau α ein signifikanter p-Wert geschätzt wurde, d.h.

$$\text{Fehler 1. Art} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{|\overline{\mathcal{A}_r}|}\sum_{a\in\overline{\mathcal{A}_r}} I\{p_{ra} < \alpha\} \tag{3.2}$$

Dabei ist $R = 200$ die Anzahl der Replikate, $\overline{\mathcal{A}_r}$ bzw. $|\overline{\mathcal{A}_r}|$ ist die Menge bzw. die Anzahl der nicht-assoziierten ROIs in Replikat r, p_{ra} ist der p-Wert zur ROI a in Replikat r, I die Indikatorfunktion und α das zuvor festgelegte Signifikanzniveau. Die Menge der nicht-assoziierten ROIs ist in allen Replikaten gleich, d.h. $\overline{\mathcal{A}_r} = \overline{\mathcal{A}}$ für alle $r = 1, \ldots, R$, und ändert sich mit dem Filterszenario (vgl. Tabelle 3.2).

3.5.2 Die empirische Teststärke

Die Teststärke eines Hypothesentests gibt an, mit welcher Wahrscheinlichkeit man sich zugunsten einer Alternativhypothese H_1 entscheidet, die wahr ist. In Bezug auf eine Simulationsstudie heißt das, dass eine krankheitsverursachende ROI mittels eines signifikanten p-Werts auch als solche erkannt wird.

Analog zum empirischen Fehler 1. Art wird für jede der 15 Gruppierungsmethoden die empirische Teststärke ermittelt. Diese ist definiert als der über alle Replikate gemittelte Anteil der assoziierten Regionen von Interesse, für die zu einem zuvor festgelegten α-Niveau ein signifikanter p-Wert geschätzt wurde, d.h.

$$\text{Empirische Teststärke} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{|\mathcal{A}_r|}\sum_{a\in\mathcal{A}_r} I\{p_{ra} < \alpha\}. \tag{3.3}$$

Dabei ist $R = 200$ die Anzahl der Replikate, $\mathcal{A}_r$ bzw. $|\mathcal{A}_r|$ st die Menge bzw. die Anzahl der assoziierten ROIs in Replikat r, p_{ra} ist der p-Wert zur ROI a in Replikat r, I die Indikatorfunktion und α das zuvor festgelegte Signifikanzniveau. Zu beachten ist, dass in dieser Simulationsstudie die Menge der assoziierten ROIs sich nur mit den Filterszenarien ändert (vgl. Tabelle 3.2), ansonsten aber in allen Replikaten gleich ist, so dass $\mathcal{A}_r = \mathcal{A}$ für alle $r = 1, \ldots, R$.

3.5.3 Die minimale Teststärke

Die minimale Teststärke ist definiert als der Anteil der Replikate, für die mindestens für eine assoziierte Region von Interesse zu einem zuvor festgelegten α-Niveau ein signifikanter p-Wert geschätzt wurde, d.h.

$$\text{Minimale Teststärke} = \frac{1}{R}\sum_{r=1}^{R} I\Big\{\Big(\sum_{a\in\mathcal{A}_r} I\{p_{ra} < \alpha\}\Big) > 0\Big\}. \tag{3.4}$$

Dabei ist $R = 200$ die Anzahl der Replikate, I die Indikatorfunktion, $\mathcal{A}_r$ die Menge der assoziierten ROIs in Replikat r, p_{ra} der p-Wert zur ROI a in Replikat r und α das zuvor festgelegte Signifikanzniveau.

Tabelle 3.2: Anzahl (#) assoziierter und nicht-assoziierter Regionen von Interesse (ROIs) für jedes der untersuchten Szenarien

Szenario	# assoz. ROIs	# nicht-assoz. ROIs
1	33	1107
2	33	1295
3	30	879
4	31	983

3.5.4 Das Signifikanzniveau α

Aufgrund der geringen Fallzahl von 697 wird in dieser Arbeit zur Beurteilung des Fehlers 1. Art und der Teststärken von einem nominellen Signifikanzniveau $\alpha = 0{,}05$ ausgegangen. Zur Auswertung der Maße wird das liberale α-Niveau-Kriterium nach Bradley (1978) verwendet, nach dem ein statistischer Test als robust gilt, falls der geschätzte Fehler 1. Art im Bereich von $[\frac{1}{2}\alpha, \frac{3}{2}\alpha]$ liegt. Das ergibt sich ein liberales Signifikanzniveau von $\frac{3}{2}\alpha = 0{,}075$.

3.6 Ergebnisse

Um die Leistung der 15 Gruppierungsmethoden zu vergleichen, werden die Ergebnisse für verschiedene Untersuchungen dargestellt: Fall-Kontroll-Status ohne Kovariablen (Abschnitt 3.6.1); mit Kovariablen (Abschnitt 3.6.2); quantitativer Phänotyp ohne Kovariablen (Abschnitt 3.6.3); und mit Kovariablen (Abschnitt 3.6.4). Hierbei werden jeweils die vier Filterszenarien betrachtet (vgl. Tabelle 3.1). Als Leistungsmaße sind jeweils der Fehler 1. Art, die empirische und die minimale Teststärke für die ROIs aus Tabelle 3.1 bestimmt worden. Die Ergebnisse sind für alle Szenarien für verschiedene MAFs und Gruppierungen tabellarisch zusammengefasst. Eine Gruppierungsmethode wird als nicht-valide angenommen, wenn sie das in Abschnitt 3.5.4 nach Bradley (1978) definierte liberale Signifikanzniveau von 0.075 nicht einhält. Ist dies für eine Gruppierungsmethode der Fall, so werden die entsprechenden geschätzten Werte der Teststärken in Klammern angegeben.

Die Quantil-Quantil-Plots in den Abbildungen 3.1a, 3.1b, 3.2a, 3.2b, 3.5a bis 3.5d, 3.7a, 3.7b, 3.8a, 3.8b und 3.11a bis 3.11d über die 200 Replikate sollen eine Bewertung der p-Wert-Verteilung der jeweiligen Gruppierungsmethoden für jedes Szenario ermöglichen. Für jede Methode sind die erwarteten gegen die beobachteten medianen zur Basis 10 log-normierten p-Werte angegeben. Assoziierte Regionen sind in Rot, nicht-assoziierte Regionen sind in Blau dargestellt. Die rote Gerade ermöglicht einen Vergleich mit dem erwarteten Zustand. Ein graues Band kennzeichnet das erste und das dritte Quartil der medianen p-Werte.

Zusätzlich wird in den Abbildungen 3.3a, 3.3b, 3.4a, 3.4b, 3.6a bis 3.6d, 3.9a, 3.9b, 3.10a, 3.10b und 3.12a bis 3.12d für jedes Szenario die These von Derkach et al. (2014) geprüft, ob die Teststärke der jeweiligen Gruppierungsmethode im Wesentlichen vom Anteil der verursachenden Varianten, dem zugrundeliegenden genetischen Effekt und der MAF abhängt. Zu diesem Zweck wird das Verhältnis verursachender Varianten zur empirischen Teststärke über die 200 Replikate betrachtet. Jeder Kreis zeigt eine assoziierte ROI. Dabei ist die Größe der Kreise proportional zur Summe der MAFs der in ihnen enthaltenen Varianten, ein Kreis ist umso heller, je größer der kumulierte Effekt aller verursachenden Varianten der jeweiligen ROI ist.

3.6.1 Binärer Phänotyp ohne Kovariablen

In Tabelle 3.3 sind der Fehler 1. Art sowie die empirische und die minimale Teststärke aller Gruppierungsmethoden für die Betrachtung des Fall-Kontroll-Status ohne Kovariablen für die ROIs angegeben, die ausschließlich aus nicht-synonymen bzw. allen Gen-Varianten mit einer MAF $\leq 0{,}05$ oder MAF $\leq 0{,}01$ bestehen.

Im oberen Teil von Tabelle 3.3 sind die Ergebnisse für ROIs mit Varianten mit einer MAF $\leq 0{,}01$ gegeben. Für die Betrachtung von ausschließlich nicht-synonymen Gen-Varianten weisen die Gruppierungsmethoden ASUM, CMC, SKAT, SKAT-O und insbesondere PWST einen erhöhten Fehler 1. Art auf, was darauf hinweist, dass diese Methoden keine validen Ergebnisse liefern. In diesem Fall sind die Kennzahlen für die Teststärken in Klammern angegeben. Bei den übrigen zehn Gruppierungsmethoden schwankt die empirische Teststärke zwischen 0,04 beim VT-Ansatz und 0,13 bei der

C-α-basierten Methode. Die Werte der minimalen Teststärke schwanken zwischen 0,7 beim VT-Ansatz und 1,00 bei CMAT.

Im Falle der Betrachtung aller Gen-Varianten eines Gens mit einer MAF $\leq$ 0,01 halten acht der 15 Methoden den Fehler 1. Art nach dem Bradley-Kriterium nicht ein. Unter diesen Methoden hat PWST mit 0,65 den größten Fehlers 1. Art. Für die gültigen Methoden, also diejenigen, die den Fehler 1.Art nach dem Bradley-Kriterium einhalten, liegt die empirische Teststärke zwischen 0,04 beim VT-Ansatz und 0,11 bei RVT2. Der Wert der minimalen Teststärke ist für die Methode RVT2 bzw. VT mit 0,85 am geringsten.

Die Ergebnisse für die ROIs mit Gen-Varianten mit einer MAF $\leq$ 0,05 befinden sich im unteren Abschnitt von Tabelle 3.3. Im Fall Gen-weiser ROI, die ausschließlich aus nicht-synonymen Varianten bestehen, halten sieben der 15 Gruppierungsmethoden den Fehler 1. Art nach dem Bradley-Kriterium ein. Von den übrigen ungültigen Methoden weist PWST mit 0,58 den höchsten Fehler 1. Art auf. Unter den gültigen Gruppierungsmethoden schwankt die empirische Teststärke zwischen 0,04 beim VT-Ansatz und 0,16 bei den Methoden CAST und RVT1; die minimale Teststärke nimmt Werte zwischen 0,71 für den VT- und 0,97 für den FPCA-Ansatz an. Lediglich ein Drittel der 15 Gruppierungsmethoden weisen einen gültigen Fehler 1. Art nach dem Bradley-Kriterium für ROIs, die alle Varianten eines Gens betrachten, auf. Hierbei liegt der kleinste Wert der empirischen Teststärke bei 0,03 im VT-Ansatz und bei 0,14 bei RVT2.

Die Abbildungen 3.1a, 3.1b, 3.2a und 3.2b sind die zugehörigen Quantil-Quantil-Plots für die oben genannten Szenarien angegeben. Für die nicht-assoziierten ROIs weichen die zur Basis 10 log-normierten medianen p-Werte bzgl. der 200 Replikate bei allen Methoden stark von den erwarteten Werten ab. Außer bei der Methode PWST, die immer zu liberale p-Werte liefert, weisen alle Methoden zu konservative p-Werte auf, das heißt der wahre p-Wert wird überschätzt. Des Weiteren sind bei den meisten Methoden sehr große Schwankungen der p-Werte für die verschiedenen Replikate für zu erkennen.

In den Abbildungen 3.3a, 3.3b, 3.4a und 3.4b sind Streudiagramme für alle Filterszenarien bei Betrachtung des Fall-Kontroll-Status ohne Kovariablen. Für die ROIs mit

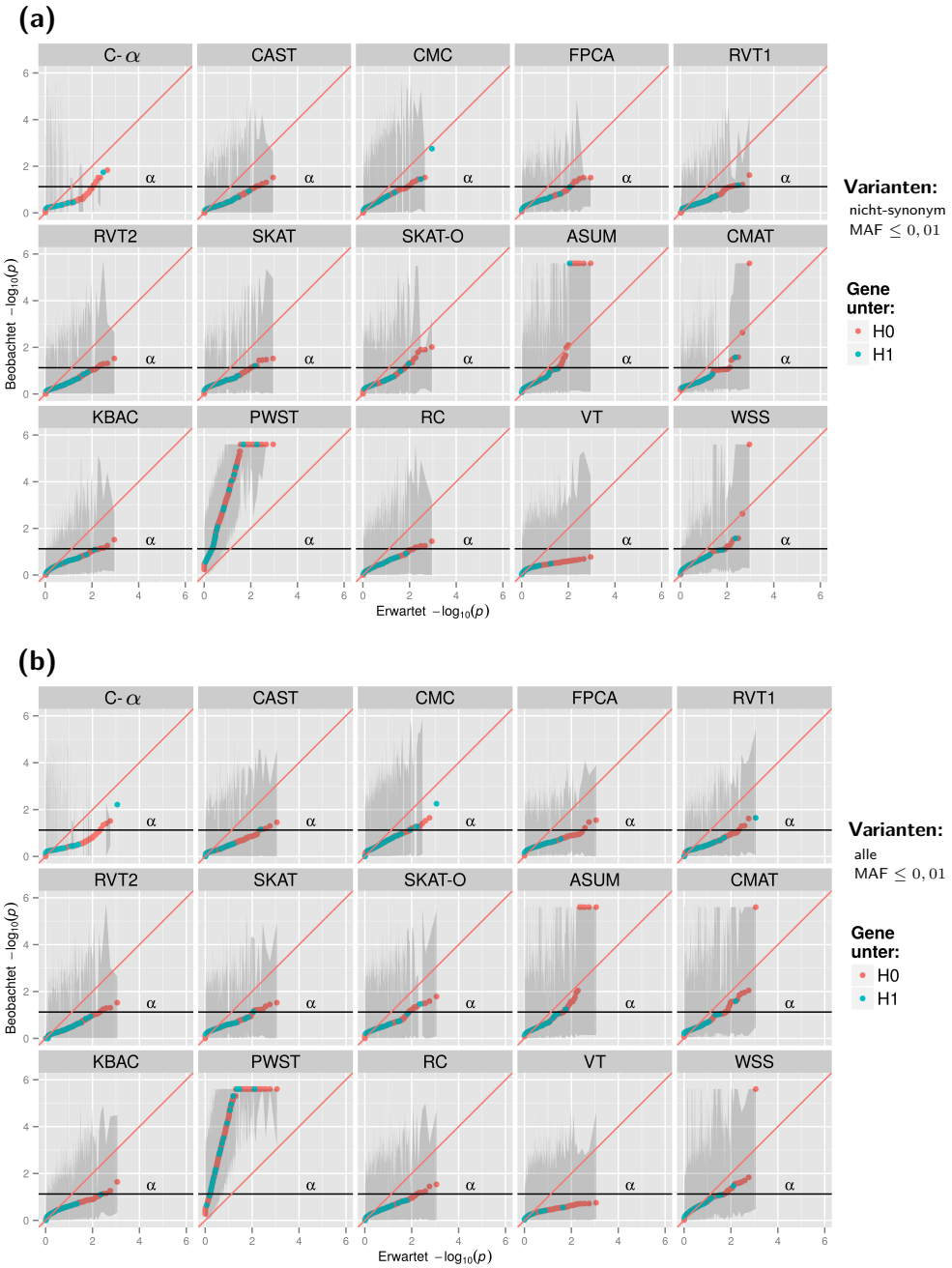


Abbildung 3.1: Quantil-Quantil-Plots der medianen zur Basis 10 log-normierten p-Werte gegen die beobachteten zur Basis 10 log-normierten p-Werte innerhalb des ersten und dritten Quartils über 200 Replikate. Betrachtung des Fall-Kontroll-Status (bei VT und PWST: quantitativer Phänotyp) ohne Kovariablen für Gen-weise Gruppierung (a) ausschließlich nicht-synonymer und (b) aller Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,01. Assoziierte Regionen sind in Rot, nicht-assoziierte in Blau dargestellt. Die rote Gerade dient dem Vergleich zum erwarteten Zustand. Das graue Band markiert das erste und dritte Quartil der medianen p-Werte.

C-α: C-alpha-based Test; CAST: cohort allelic sum test; CMC: combined multivariate cluster; FPCA: functional principal component analysis; RVT: rare variant test 1 und 2; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT; ASUM: adaptive summation; CMAT: cumulative minor-allele test; KBAC: kernel-based adaptive cluster; PWST: p-value weighted sum test; RC: Rarecover; VT: variable threshold; WSS: weighted sum statistic.

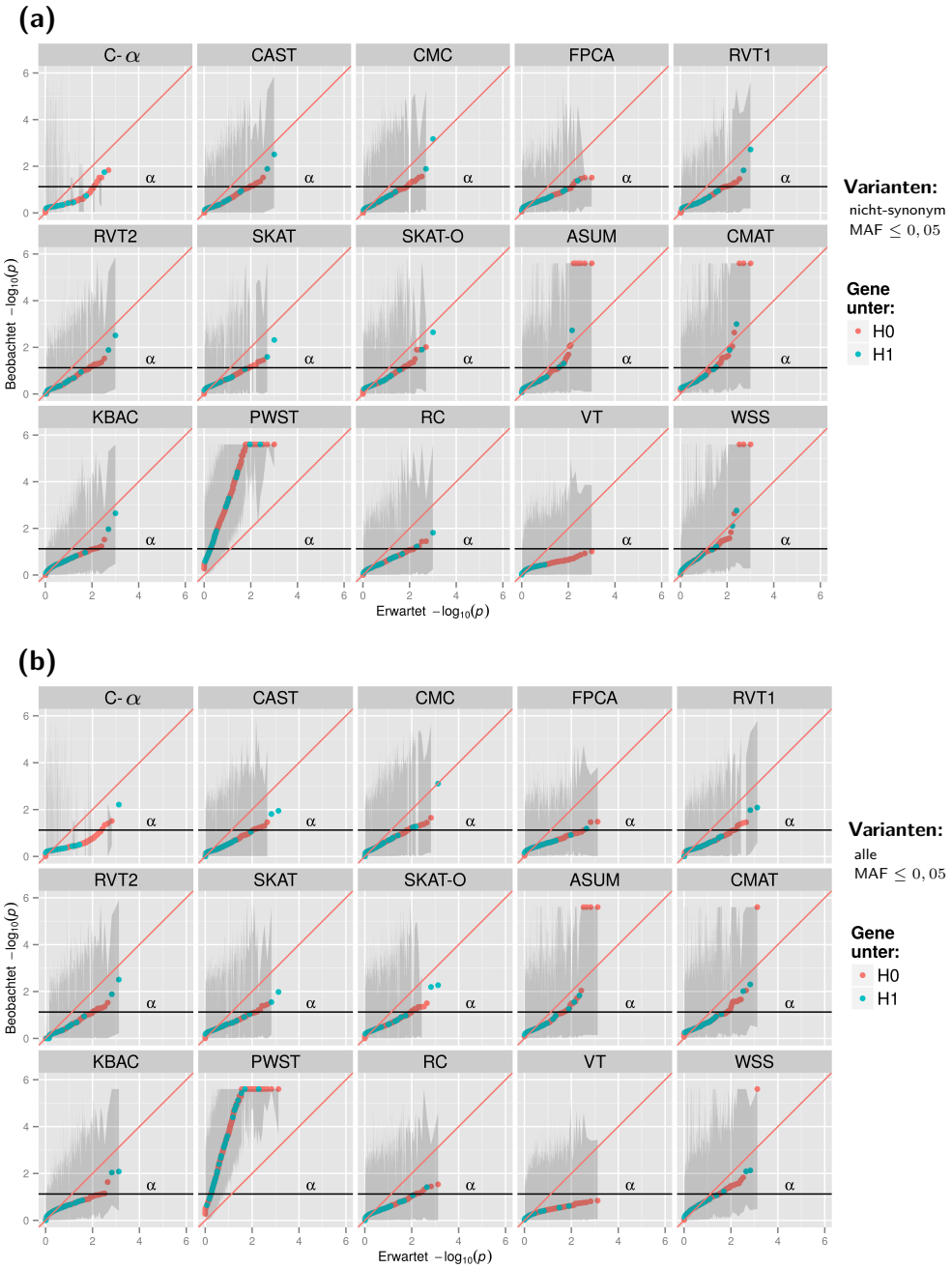


Abbildung 3.2: Quantil-Quantil-Plots der medianen zur Basis 10 log-normierten p-Werte gegen die beobachteten zur Basis 10 log-normierten p-Werte innerhalb des ersten und dritten Quartils über 200 Replikate. Betrachtung des Fall-Kontroll-Status (bei VT und PWST: quantitativer Phänotyp) ohne Kovariablen für Gen-weise Gruppierung (a) ausschließlich nicht-synonymer und (b) aller Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,05. Assoziierte Regionen sind in Rot, nicht-assoziierte in Blau dargestellt. Die rote Gerade dient dem Vergleich zum erwarteten Zustand. Das graue Band markiert das erste und dritte Quartil der medianen p-Werte.

C-α: C-alpha-based Test; CAST: cohort allelic sum test; CMC: combined multivariate cluster; FPCA: functional principal component analysis; RVT: rare variant test 1 und 2; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT; ASUM: adaptive summation; CMAT: cumulative minor-allele test; KBAC: kernel-based adaptive cluster; PWST: p-value weighted sum test; RC: Rarecover; VT: variable threshold; WSS: weighted sum statistic.

Tabelle 3.3: Teststärke und Fehler 1. Art bei Gen-weiser Gruppierung aller bzw. ausschließlich nicht-synonymer Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,01 bzw. $\leq$ 0,05. Betrachtet wird der Fall-Kontroll-Status ohne Kovariablen. Der Fehler 1. Art, die empirische und die minimale Teststärke wurden über 200 Replikate gemittelt. Bei einem erhöhten Fehler 1. Art sind die Werte der Teststärke als nicht valide in Klammern angegeben.

		nicht-synonyme Varianten			alle Varianten		
MAF	Methode	Fehler 1.Art	Empirische Teststärke	Minimale Teststärke	Fehler 1.Art	Empirische Teststärke	Minimale Teststärke
$\leq$ 0,01	ASUM	0,12	(0,19)	(1.00)	0,12	(0,15)	(1,00)
	C-α	0,07	0,13	0,90	0,06	0,10	0,91
	CAST	0,06	0,11	0,93	0,06	0,09	0,97
	CMAT	0,05	0,10	1,00	0,12	(0,16)	(1,00)
	CMC	0,10	(0,16)	(0,93)	0,11	(0,20)	(0,97)
	FPCA	0,06	0,11	0,93	0,06	0,08	0,97
	KBAC	0,03	0,06	0,77	0,04	0,06	0,88
	PWST	0,50	(0,58)	(1,00)	0,65	(0,85)	(1,00)
	RC	0,07	0,12	0,93	0,08	(0,13)	(0,94)
	RVT1	0,05	0,11	0,77	0,06	0,10	0,91
	RVT2	0,06	0,12	0,93	0,05	0,11	0,85
	SKAT	0,08	(0,12)	(0,93)	0,08	(0,11)	(0,97)
	SKAT-O	0,10	(0,16)	(0,93)	0,10	(0,14)	(1,00)
	VT	0,07	0,04	0,70	0,06	0,04	0,85
	WSS	0,05	0,11	0,90	0,12	(0,17)	(0,94)
$\leq$ 0,05	ASUM	0,12	(0,21)	(1.00)	0,12	(0,17)	(1.00)
	C-α	0,07	0,14	0,90	0,06	0,10	0,91
	CAST	0,07	0,16	0,94	0,08	(0,14)	(0,97)
	CMAT	0,13	(0,23)	(1,00)	0,12	(0,19)	(1,00)
	CMC	0,10	(0,16)	(0,94)	0,11	(0,20)	(0,97)
	FPCA	0,07	0,14	0,97	0,07	0,10	0,97
	KBAC	0,05	0,12	0,87	0,05	0,11	0,91
	PWST	0,58	(0,68)	(1,00)	0,61	(0,84)	(1,00)
	RC	0,08	(0,15)	(0,94)	0,09	(0,16)	(0,94)
	RVT1	0,07	0,16	0,84	0,08	(0,16)	(0,97)
	RVT2	0,07	0,15	0,90	0,06	0,14	0,85
	SKAT	0,09	(0,16)	(0,94)	0,10	(0,16)	(0,97)
	SKAT-O	0,11	(0,19)	(0,94)	0,11	(0,19)	(1,00)
	VT	0,07	0,04	0,71	0,06	0,03	0,76
	WSS	0,13	(0,25)	(0,97)	0,14	(0,22)	(0,94)

ASUM: adaptive summation; C-α: C-alpha-based Test; CAST: cohort allelic sum test; CMAT: cumulative minor-allele test; CMC: combined multivariate cluster; FPCA: functional principal component analysis; KBAC: kernel-based adaptive cluster; PWST: p-value weighted sum test; RC: Rarecover; RVT: rare variant test 1 und 2; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT; VT: variable threshold; WSS: weighted sum statistic.

ausschließlich nicht-synonymen Varianten und einer MAF $\leq$ 0,01 (Abbildung 3.3a) haben lediglich die Methoden C-α und CMC bei einer ROI eine Teststärke größer 0,80, wobei C-α als einzige Methode den Fehler 1. Art nach dem Bradley-Kriterium einhält. Außer bei der Methode PWST ist in diesem Szenario keine klare Tendenz hinsichtlich der kumulierten MAF, des kumulierten Effekts oder des Anteils verursachender Varianten und der entsprechenden Teststärke zu erkennen. Dabei entspricht die kumulierte MAF bzw. der kumulierte Effekt, der Summe der MAFs bzw. Effektstärken der Varianten in der ROI. Da PWST den Fehler 1. Art deutlich überschreitet ist dieses Ergebnis nicht valide. In dem Szenario mit allen Gen-Varianten mit einer MAF $\leq$ 0,01 (Abbildung 3.3b) weisen viele der Methoden schon bei einem Anteil von ca. 40% verursachender Varianten eine höhere Teststärke auf, wenn gleichzeitig die kumulierte MAF $>$ 0,10 und der kumulierte genetische Effekt $>$ 5 für dieselbe ROI ist. Ist die kumulierte MAF allerdings kleiner als 0,05, werden von validen Methoden selbst bei hohen kumulierten Effekten von ca. 7 und hohen Anteilen verursachender Varianten kaum Teststärken von mehr als 0,60 erreicht. In den Szenarien von ROIs mit Varianten mit einer MAF $\leq$ 0,05 wird die Wirkung der Höhe der Effektstärken, der kumulierten MAF und des Anteils verursachender Varianten auf die Teststärke weiter deutlich. Selbst bei einem vergleichsweise geringen Anteil verursachender Varianten von ca. 30–40% kann bei gleichzeitig hohen kumulierten MAFs und Effekten eine Teststärke von mehr als 0,80 unter den validen Methoden erreicht werden.

3.6.2 Binärer Phänotyp mit Kovariablen

In Tabelle 3.4 sind der Fehler 1. Art und die Teststärken bei Betrachtung des Fall-Kontroll-Status mit Kovariablen für die ROIs bestehend aus entweder nicht-synonymen oder aus allen Gen-Varianten mit einer MAF $\leq$ 0,01 und $\leq$ 0,05 angegeben. Gegeben sind die Ergebnisse für die drei Methoden RVT1, SKAT und SKAT-O, die eine Einbeziehung von Kovariablen ermöglichen. Nach dem Bradley-Kriterium liefern die Methoden SKAT und SKAT-O für alle Fälle von ROIs einen erhöhten Fehler 1. Art. Dabei steigt der Fehler sowohl mit steigender MAF als auch mit der Erweiterung von ausschließlich nicht-synonymen auf alle Varianten eines Gens. Die Methode RVT1 liefert bis auf das Szenario mit ROIs mit allen Varianten eines Gens und einer MAF $\leq$ 0,05 keinen erhöhten Fehler 1. Art, wobei auch hier der Wert des Fehlers mit der MAF und der Art der Varianten steigt. Analog steigt auch die

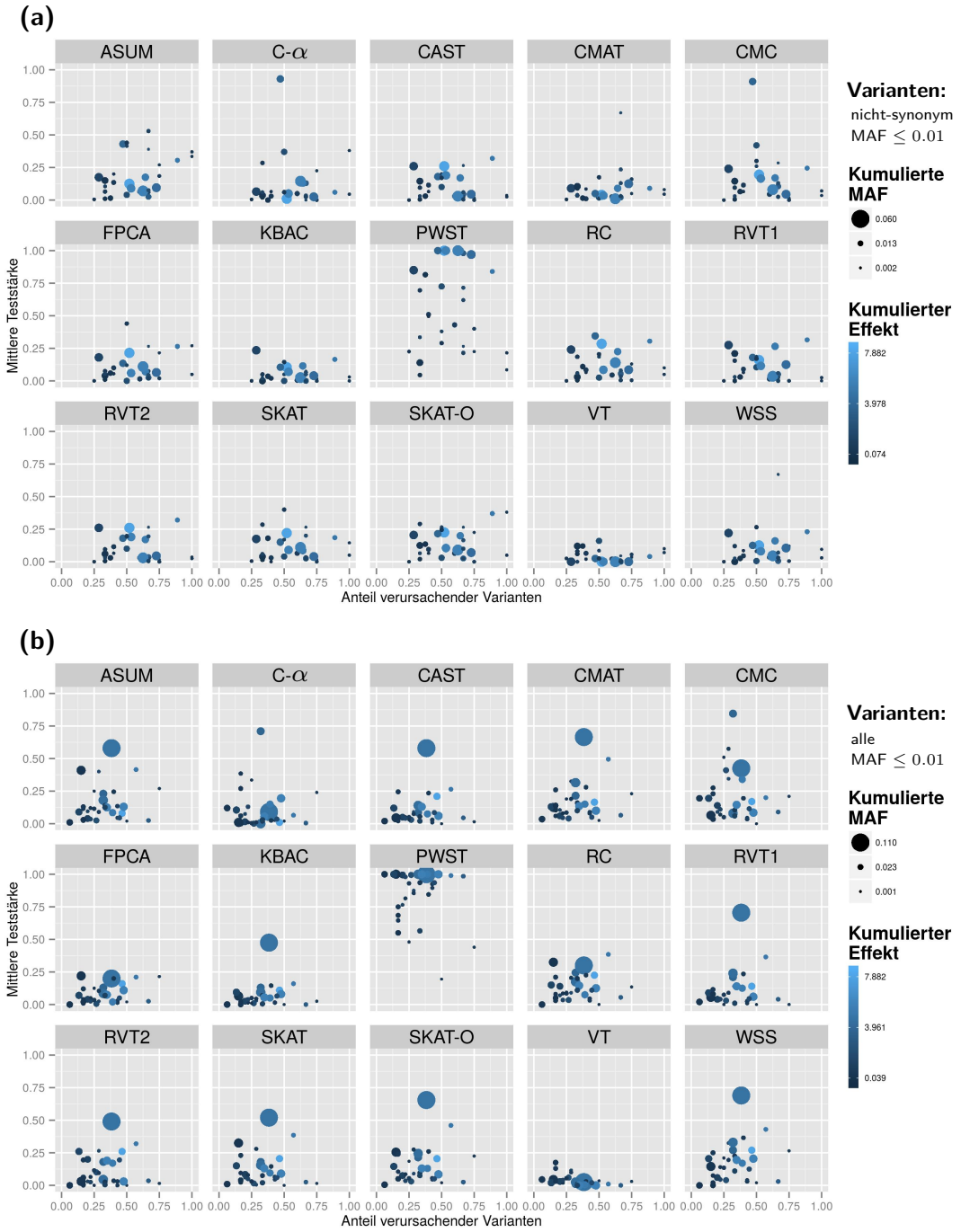


Abbildung 3.3: Streudiagramme der assoziierten ROIs bzgl. des Anteils der verursachenden Varianten und der empirischen Teststärke über 200 Replikate. Die Größe der Punkte ist proportional zur Summe der MAFs, je heller die Punkte, desto größer ist der aufsummierte Effekt der Varianten der jeweiligen ROI. Betrachtung des Fall-Kontroll-Status (bei VT und PWST: quantitativer Phänotyp) ohne Kovariablen bei Gen-weiser Gruppierung (a) ausschließlich nicht-synonymer und (b) aller Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq 0{,}01$.
ASUM: adaptive summation; C-α: C-alpha-based Test; CAST: cohort allelic sum test; CMAT: cumulative minor-allele test; CMC: combined multivariate cluster; FPCA: functional principal component analysis; KBAC: kernel-based adaptive cluster; PWST: p-value weighted sum test; RC: Rarecover; RVT: rare variant test 1 und 2; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT; VT: variable threshold; WSS: weighted sum statistic.

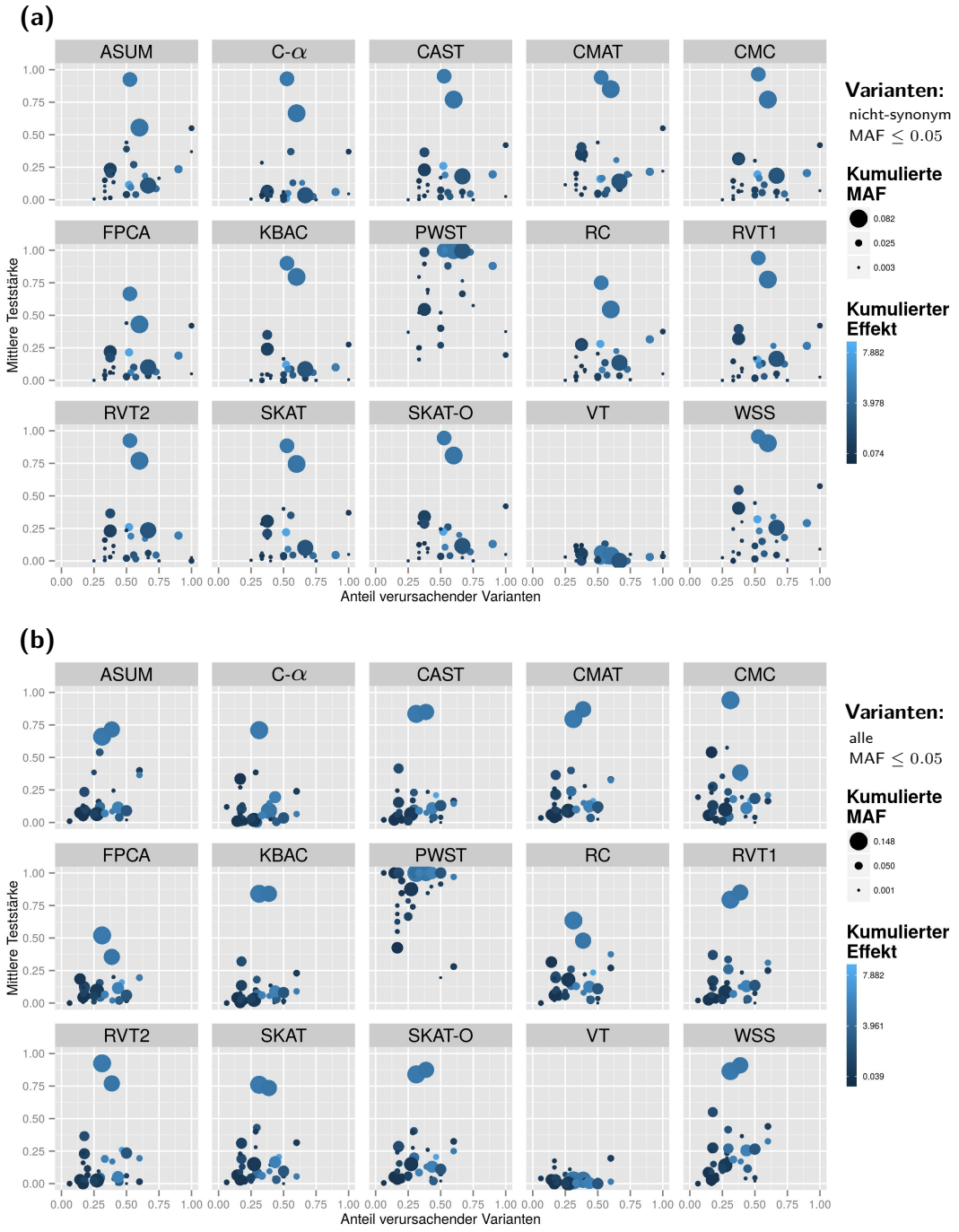


Abbildung 3.4: Streudiagramme der assoziierten ROIs bzgl. des Anteils der verursachenden Varianten und der empirischen Teststärke über 200 Replikate. Die Größe der Punkte ist proportional zur Summe der MAFs, je heller die Punkte, desto größer ist der aufsummierte Effekt der Varianten der jeweiligen ROI. Betrachtung des Fall-Kontroll-Status (bei VT und PWST: quantitativer Phänotyp) ohne Kovariablen bei Gen-weiser Gruppierung (a) nur nicht-synonymer und (b) aller Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,05.

ASUM: adaptive summation; C-α: C-alpha-based Test; CAST: cohort allelic sum test; CMAT: cumulative minor-allele test; CMC: combined multivariate cluster; FPCA: functional principal component analysis; KBAC: kernel-based adaptive cluster; PWST: p-value weighted sum test; RC: Rarecover; RVT: rare variant test 1 und 2; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT; VT: variable threshold; WSS: weighted sum statistic.

empirische Teststärke. Der größte gültige Wert der empirischen Teststärke ist 0,16 während der kleinste 0,10 ist. Die gültigen Werte der minimalen Teststärke liegen zwischen 0,8 bei der Gruppierung von nicht-synonymen Varianten eines Gens mit einer MAF $\leq$ 0,01 und 0,97 bei der Gruppierung von allen Gen-Varianten mit der gleichen MAF-Grenze. Dies ist kontra-intuitiv, da bei der Betrachtung des gesamten Gens gegenüber dem mit ausschließlich nicht-synonymen Varianten nur neutrale Varianten hinzukommen.

Tabelle 3.4: Teststärke und Fehler 1. Art bei Gen-weiser Gruppierung aller bzw. ausschließlich nicht-synonymer Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,01 bzw. $\leq$ 0,05. Betrachtet wird der Fall-Kontroll-Status mit Kovariablen. Der Fehler 1. Art, die empirische und die minimale Teststärke wurden über 200 Replikate gemittelt. Bei einem erhöhten Fehler 1. Art sind die Werte der Teststärke als nicht-valide, in Klammern angegeben.

		nicht-synonyme Varianten			alle Varianten		
MAF	Methode	Fehler 1.Art	Empirische Teststärke	Minimale Teststärke	Fehler 1.Art	Empirische Teststärke	Minimale Teststärke
$\leq$ 0,01	RVT1	0,05	0,10	0,80	0,06	0,11	0,97
	SKAT	0,08	(0,12)	(1,00)	0,08	(0,11)	(1,00)
	SKAT-O	0,10	(0,17)	(0,97)	0,10	(0,15)	(1,00)
$\leq$ 0,05	RVT1	0,07	0,16	0,87	0,08	(0,17)	(0,97)
	SKAT	0,09	(0,17)	(1,00)	0,10	(0,16)	(1,00)
	SKAT-O	0,11	(0,20)	(0,97)	0,11	(0,19)	(1,00)

RVT1: rare variant test 1; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT.

Die entsprechenden Quantil-Quantil-Plots in Abbildungen 3.5a bis 3.5d zeigen, dass die p-Werte aller in diesem Szenario betrachteten Gruppierungsmethoden im Median über die 200 Replikate nicht den erwarteten Werten entsprechen. Dabei ist die Varianz der beobachteten p-Werte über die 200 Replikate sehr groß und steigt entgegengesetzt proportional zur Größe der p-Werte.

Die zugehörigen Streudiagramme sind in den Abbildungen 3.6a bis 3.6d zu finden. Im Gruppierungsszenario mit ausschließlich nicht-synonymen Varianten und einer MAF $\leq$ 0,01 ist für keine der Methoden ein eindeutiger Zusammenhang zwischen dem Anteil verursachender Varianten und der Höhe der empirischen Teststärke zu erkennen. In den übrigen drei Gruppierungsszenarien wird, wie auch schon bei der Betrachtung des Fall-Kontroll-Status ohne Kovariablen, deutlich, dass mit dem

gleichzeitigen Ansteigen von kumulierter MAF und kumuliertem Effekt auch die Teststärke größer wird. Nimmt der Wert nur einer dieser beiden kumulierten Größen zu, führt ein größerer Anteil verursachender Varianten nicht unbedingt zu höheren Teststärken.

3.6.3 Quantitativer Phänotyp ohne Kovariablen

In Tabelle 3.5 sind der Fehler 1. Art sowie die Szenarien-relevanten Teststärken aller Gruppierungsmethoden für die Betrachtung des quantitativen Phänotyps ohne Kovariablen für die ROIs bestehend aus entweder nicht-synonymen oder allen Gen-Varianten mit einer MAF $\leq 0{,}01$ bzw. $\leq 0{,}05$ angegeben. In beiden Szenarien ist zu sehen, dass von den betrachteten Gruppierungsmethoden alle (mit Ausnahme des VT-Ansatzes) unabhängig von der MAF einen erhöhten Fehler 1. Art aufweisen. Unter den nicht-validen Methoden erreicht PWST unabhängig von der Art der Varianten und der entsprechenden MAF stets die höchsten Werte zwischen 0,50 und 0,65. Die übrigen nicht-validen Methoden weisen Fehler zwischen 0,09 und 0,10 auf. Der für alle Gruppierungsszenarien gültige Ansatz VT erreicht unabhängig von der MAF einen Fehler 1. Art von 0,06 bei der Gruppierung aller Varianten und einen Wert von 0,07 bei der Gruppierung ausschließlich nicht-synonymer Varianten eines Gens. Bei der Gruppierung aller Varianten mit einer MAF $\leq 0{,}05$ liegt die empirische Teststärke des VT-Ansatzes bei 0,03, bei den übrigen Gruppierungsszenarien beträgt sie 0,04. Die minimale Teststärke des VT-Ansatzes schwankt zwischen 0,70 bei der Gruppierung ausschließlich nicht-synonymer Varianten mit einer MAF $\leq 0{,}01$ und 0,85 bei der Gruppierung aller Varianten eines Gens mit einer MAF $\leq 0{,}01$.

In Abbildungen 3.7a, 3.7b, 3.8a und 3.8b sind die zugehörigen Quantil-Quantil-Plots für die oben genannten Szenarien gegeben. Auch in diesem Szenario ist zu sehen, dass die medianen p-Werte der nicht-assoziierten ROIs für alle der hier betrachteten Gruppierungsmethoden stark von den erwarteten Werten abweichen. Abgesehen von der Methode PWST, die ausschließlich zu liberale p-Werte liefert, weisen die übrigen Methoden zu konservative p-Werte auf. Weiterhin ist für die meisten Methoden eine sehr große Schwankung der p-Werte innerhalb der Replikate zu erkennen. Die größte Schwankung unter den betrachteten Gruppierungsansätzen weisen dabei RVT2 und SKAT auf, unabhängig vom Gruppierungsszenario.

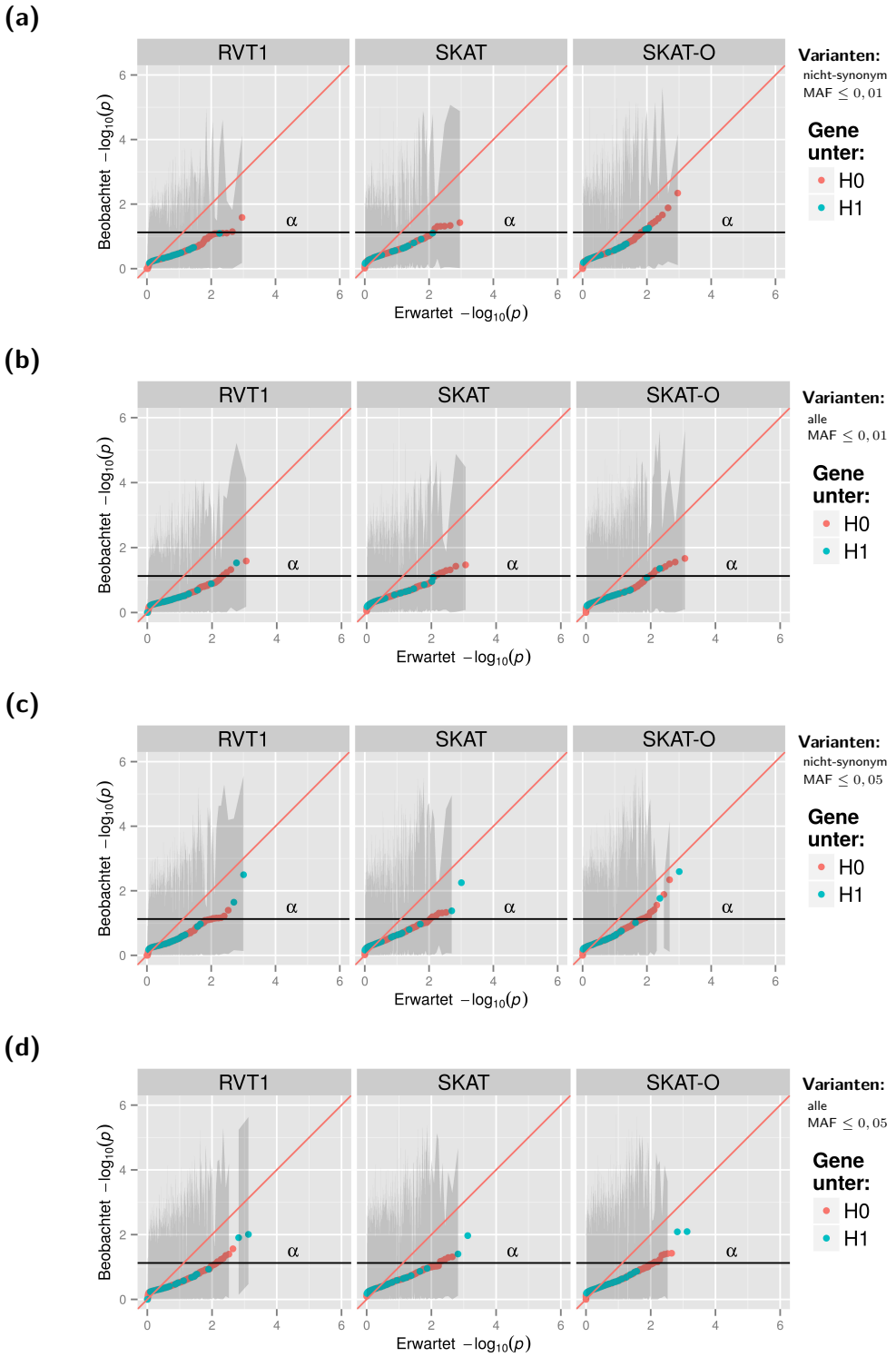


Abbildung 3.5: Quantil-Quantil-Plots der medianen zur Basis 10 log-normierten p-Werte gegen die beobachteten zur Basis 10 log-normierten p-Werte innerhalb des ersten und dritten Quartils über 200 Replikate. Betrachtung des Fall-Kontroll-Status mit Kovariablen bei Gen-weiser Gruppierung (a) ausschließlich nicht-synonymer und (b) aller Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq 0{,}01$ und Gen-weise Gruppierung (c) ausschließlich nicht-synonymer und (d) aller Varianten mit einer MAF $\leq 0{,}05$

RVT1: rare variant test 1; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT.

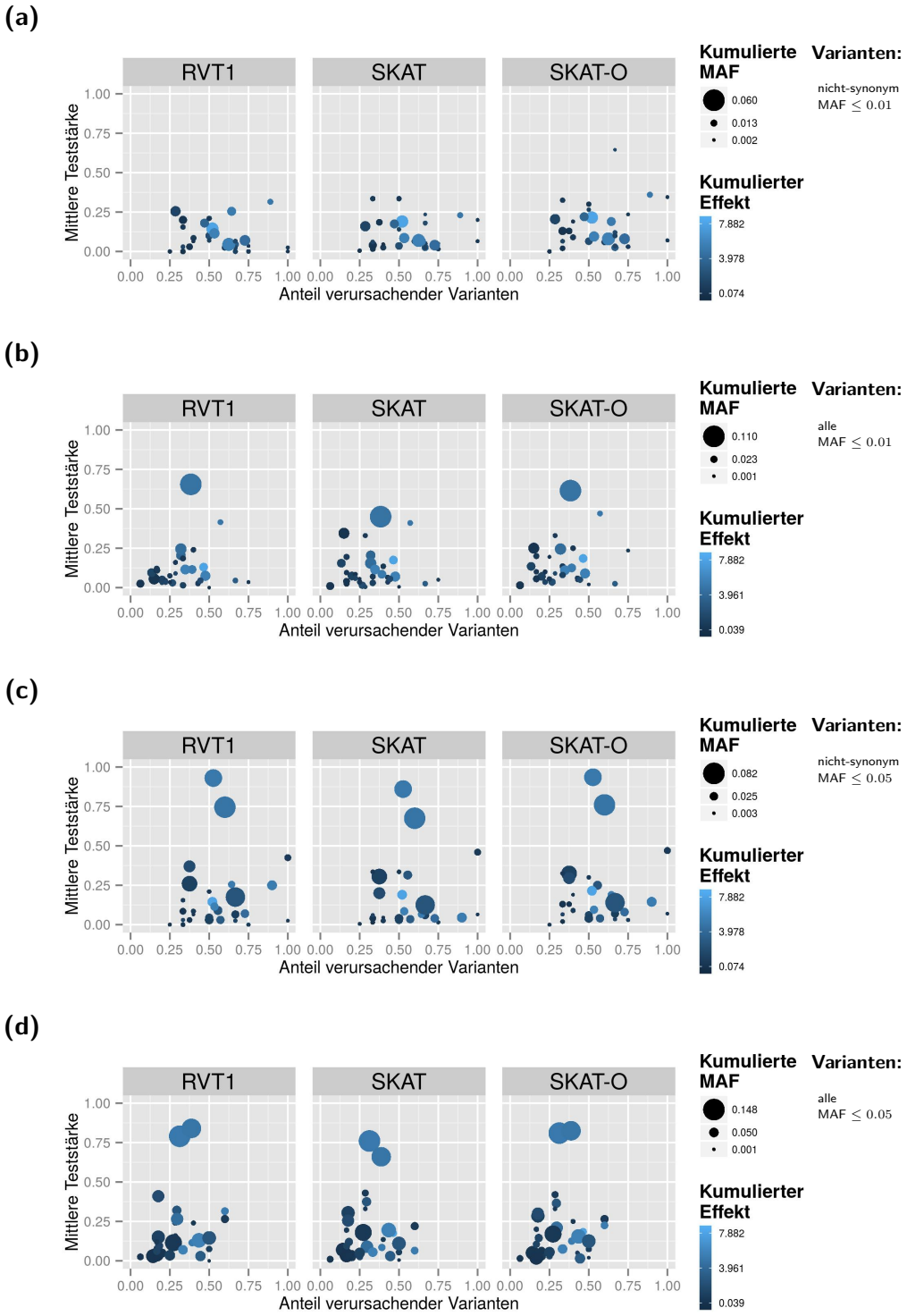
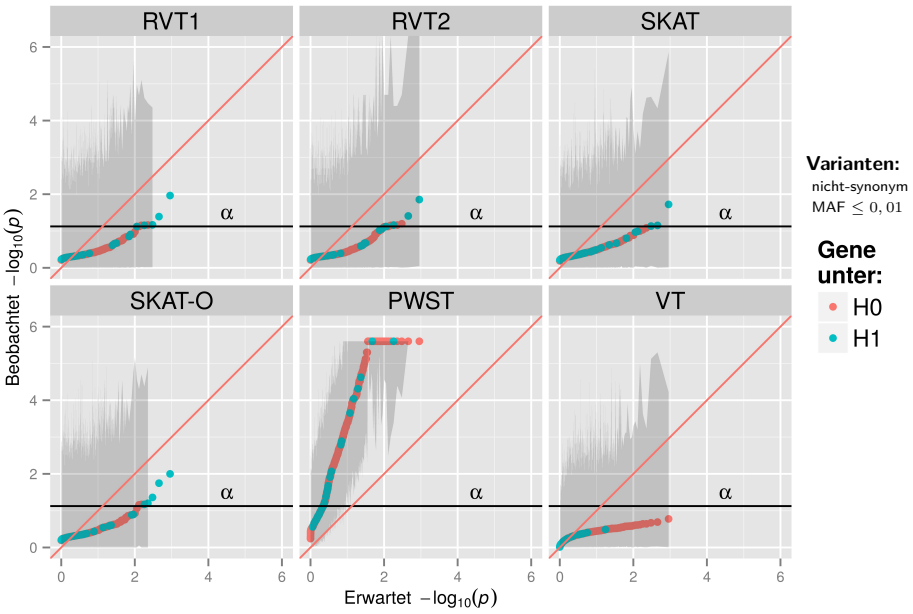


Abbildung 3.6: Streudiagramme der assoziierten ROIs bzgl. des Anteils der verursachenden Varianten und der empirischen Teststärke über 200 Replikate. Die Größe der Punkte ist proportional zur Summe der MAFs, je heller die Punkte, desto größer ist der aufsummierte Effekt der Varianten der jeweiligen ROI. Betrachtung des Fall-Kontroll-Status mit Kovariablen bei Gen-weiser Gruppierung (a) ausschließlich nicht-synonymer und (b) aller Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,01 und Gen-weiser Gruppierung (c) ausschließlich nicht-synonymer und (d) aller Varianten mit einer MAF $\leq$ 0,05.

RVT1: rare variant test 1; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT.

(a)



(b)

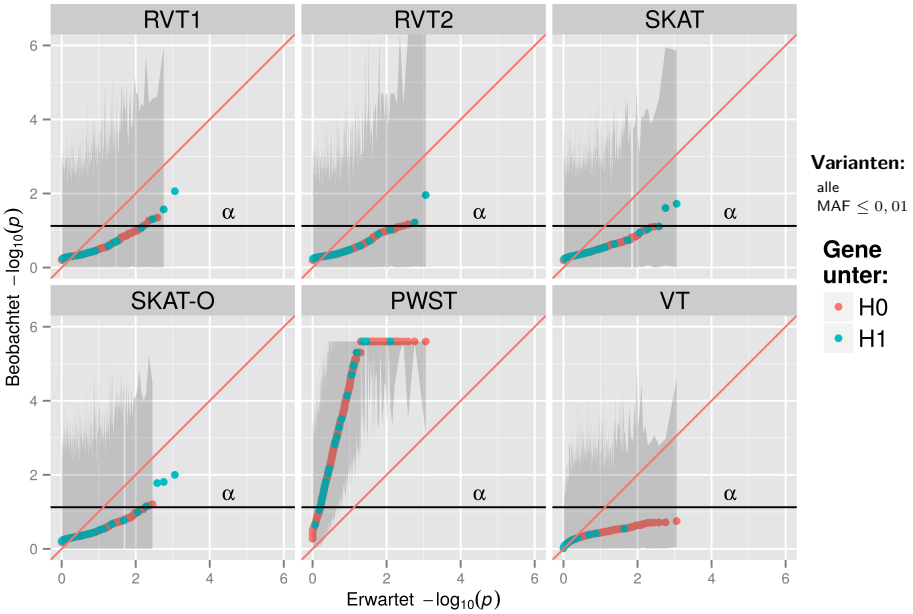
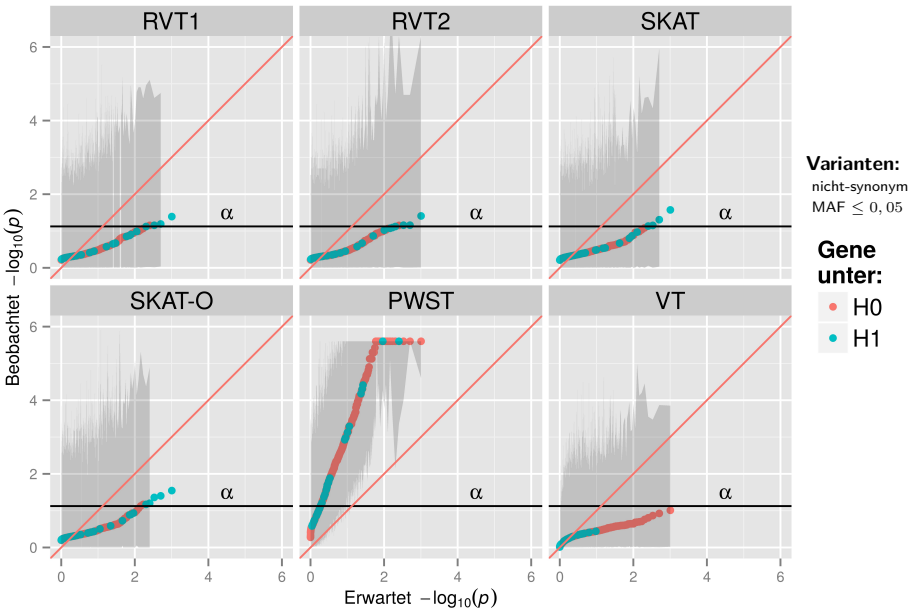


Abbildung 3.7: Quantil-Quantil-Plots der medianen zur Basis 10 log-normierten p-Werte gegen die beobachteten zur Basis 10 log-normierten p-Werte innerhalb des ersten und dritten Quartils über 200 Replikate. Betrachtung des quantitativen Phänotyps ohne Kovariablen bei Gen-weiser Gruppierung (a) ausschließlich nicht-synonymer und (b) aller Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,01. Assoziierte Regionen sind in Rot, nicht-assoziierte in Blau dargestellt. Die rote Gerade dient dem Vergleich zum erwarteten Zustand. Das graue Band markiert das erste und dritte Quartil der medianen p-Werte.

RVT: rare variant test 1 und 2; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT; PWST: p-value weighted sum test; VT: variable threshold.

(a)



(b)

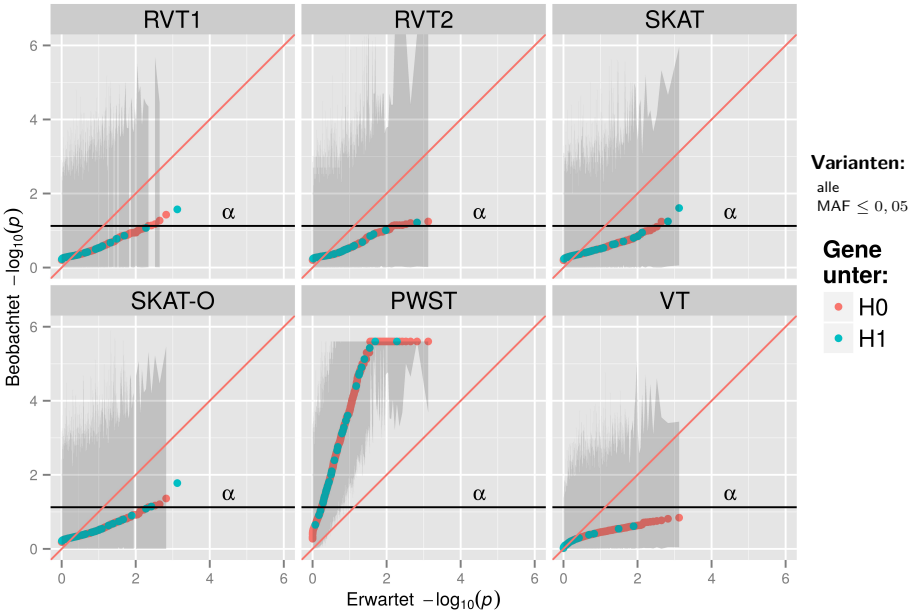


Abbildung 3.8: Quantil-Quantil-Plots der medianen zur Basis 10 log-normierten p-Werte gegen die beobachteten zur Basis 10 log-normierten p-Werte innerhalb des ersten und dritten Quartils über 200 Replikate. Betrachtung des quantitativen Phänotyps ohne Kovariablen bei Gen-weiser Gruppierung (a) ausschließlich nicht-synonymer und (b) aller Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,05. Assoziierte Regionen sind in Rot, nicht-assoziierte in Blau dargestellt. Die rote Gerade dient dem Vergleich zum erwarteten Zustand. Das graue Band markiert das erste und dritte Quartil der medianen p-Werte..

RVT: rare variant test 1 und 2; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT; PWST: p-value weighted sum test; VT: variable threshold.

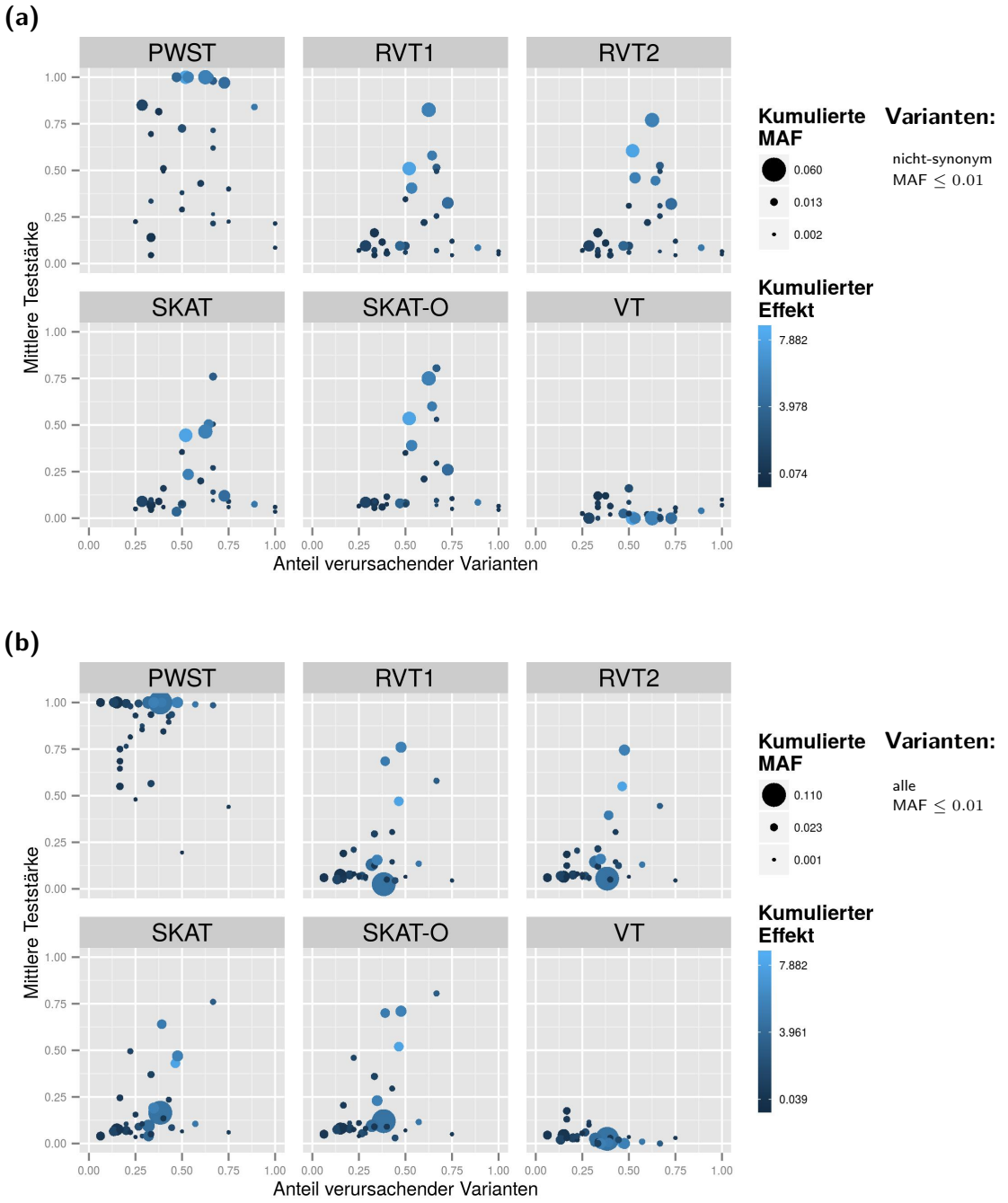


Abbildung 3.9: Streudiagramme der assoziierten ROIs bzgl. des Anteils der verursachenden Varianten und der empirischen Teststärke über 200 Replikate. Die Größe der Punkte ist proportional zur Summe der MAFs, je heller die Punkte, desto größer ist der aufsummierte Effekt der Varianten der jeweiligen ROI. Betrachtung des quantitativen Phänotyps ohne Kovariablen bei Gen-weiser Gruppierung (a) ausschließlich nicht-synonymer und (b) aller Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,01.

PWST: p-value weighted sum test; RVT: rare variant test 1 und 2; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT; VT: variable threshold.

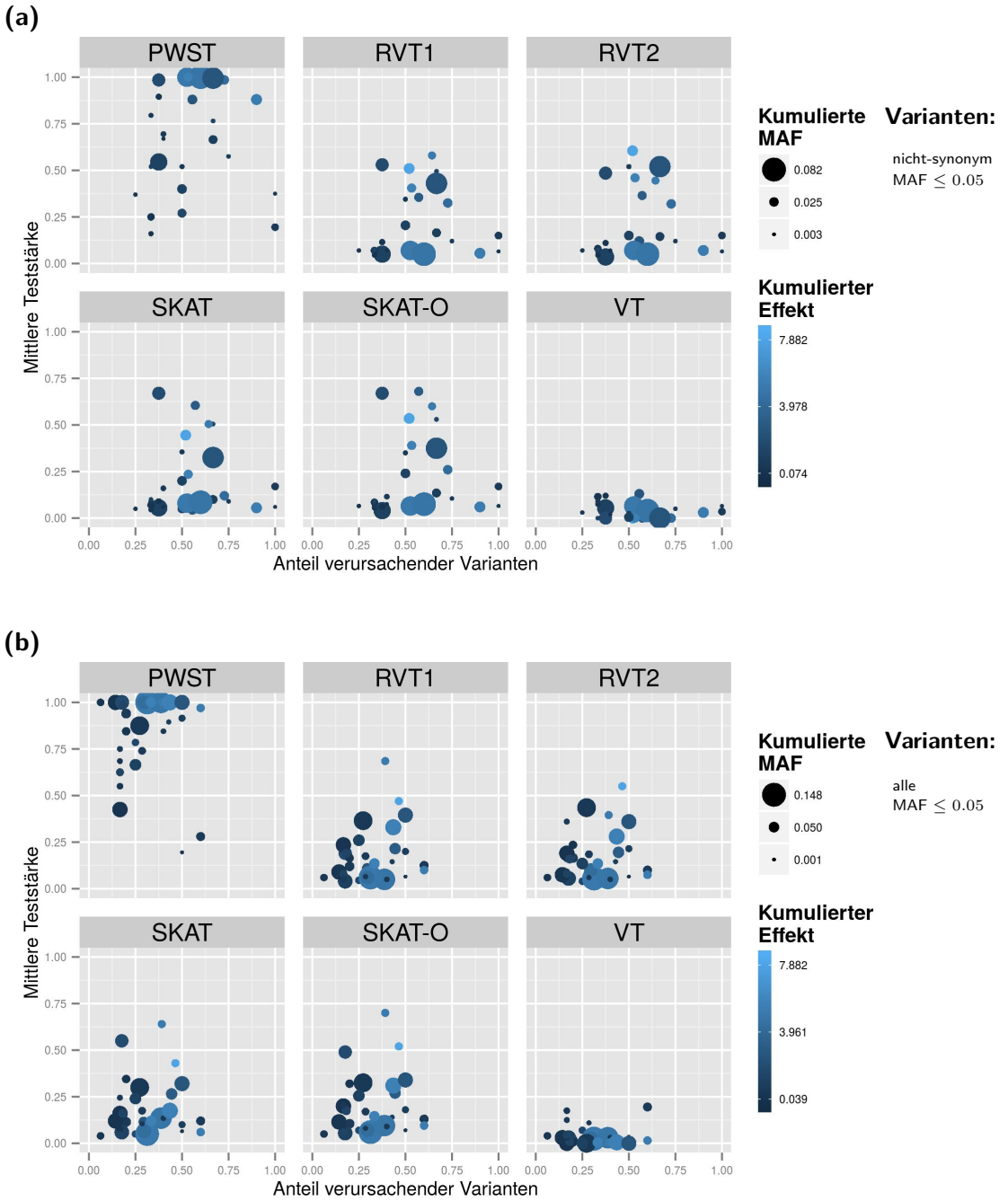


Abbildung 3.10: Streudiagramme der assoziierten ROIs bzgl. des Anteils der verursachenden Varianten und der empirischen Teststärke über 200 Replikate. Die Größe der Punkte ist proportional zur Summe der MAFs, je heller die Punkte, desto größer ist der aufsummierte Effekt der Varianten der jeweiligen ROI. Betrachtung des quantitativen Phänotyps ohne Kovariablen bei Gen-weiser Gruppierung (a) ausschließlich nicht-synonymer und (b) aller Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,05.

PWST: p-value weighted sum test; RVT: rare variant test 1 und 2; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT; VT: variable threshold.

Tabelle 3.5: Teststärke und Fehler 1. Art bei Gen-weiser Gruppierung aller bzw. ausschließlich nicht-synonymer Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq 0{,}01$ bzw. $\leq 0{,}05$. Betrachtet wird der quantitative Phänotyp ohne Kovariablen. Der Fehler 1. Art, die empirische und die minimale Teststärke wurden über 200 Replikate gemittelt. Bei einem erhöhten Fehler 1. Art sind die Werte der Teststärke als nicht-valide in Klammern angegeben.

		nicht-synonyme Varianten			alle Varianten		
MAF	Methode	Fehler 1.Art	Empirische Teststärke	Minimale Teststärke	Fehler 1.Art	Empirische Teststärke	Minimale Teststärke
$\leq 0{,}01$	PWST	0,50	(0,58)	(1,00)	0,65	(0,85)	(1,00)
	RVT1	0,09	(0,20)	(1,00)	0,10	(0,17)	(1,00)
	RVT2	0,09	(0,21)	(1,00)	0,10	(0,16)	(1,00)
	SKAT	0,09	(0,18)	(1,00)	0,10	(0,18)	(1,00)
	SKAT-O	0,10	(0,21)	(1,00)	0,10	(0,19)	(1,00)
	VT	0,07	0,04	0,70	0,06	0,04	0,85
$\leq 0{,}05$	PWST	0,58	(0,68)	(1,00)	0,61	(0,84)	(1,00)
	RVT1	0,10	(0,19)	(1,00)	0,11	(0,17)	(1,00)
	RVT2	0,10	(0,20)	(1,00)	0,11	(0,17)	(1,00)
	SKAT	0,09	(0,18)	(1,00)	0,10	(0,18)	(1,00)
	SKAT-O	0,10	(0,21)	(1,00)	0,11	(0,19)	(1,00)
	VT	0,07	0,04	0,71	0,06	0,03	0,76

PWST: p-value weighted sum test; RVT: rare variant test 1 und 2; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT; VT: variable threshold.

In den Abbildungen 3.9a, 3.9b, 3.10a und 3.10b sind die Streudiagramme für die entsprechenden Gruppierungsmethoden in allen untersuchten Gruppierungsszenarien zu finden. Der einzige valide Ansatz (VT) zeigt in keinem Gruppierungsszenario zu irgendeinem der hier betrachten möglichen Einflussfaktoren (Anzahl verursachender Varianten, kumulierte MAF, kumulierter Effekt) einen deutlichen Zusammenhang zur gemittelten Teststärke. Für die Gruppierungsszenarien ist jeweils in den übrigen nicht-validen Gruppierungsmethoden (außer PWST) zu sehen, dass die einzelnen assoziierten ROIs einen ähnlichen Zusammenhang zwischen dem Anteil verursachender Varianten und der empirischen Teststärke aufweisen. Für die Szenarien mit MAF $\leq 0{,}01$ ist ein additiver Effekt der drei Variablen, Anzahl verursachender Varianten, kumulierte MAF und kumulierter Effekt erkennbar. Sind alle drei Variablen hinreichend groß, werden empirische Teststärken von über 0,75 erreicht.

3.6.4 Quantitativer Phänotyp mit Kovariablen

In Tabelle 3.6 sind der Fehler 1. Art sowie die empirische und die minimale Teststärke bei Betrachtung des quantitativen Phänotyps mit Kovariablen für die Gruppierung ausschließlich nicht-synonymer bzw. aller Varianten eines Gens mit MAF $\leq$ 0,01 bzw. $\leq$ 0,05 angegeben. Nur die drei Gruppierungsmethoden RVT1, SKAT und SKAT-O ermöglichen eine Einbeziehung von Kovariablen und lagen in einer entsprechenden Software vor. Nach dem Bradley-Kriterium liefert keine der betrachteten Methoden in einem der vier Gruppierungsszenarien zuverlässige Ergebnisse, da der Fehler 1. Art jeweils zu hoch ist. Dieser schwankt zwischen 0,09 und 0,11.

Tabelle 3.6: Teststärke und Fehler 1. Art bei Gen-weiser Gruppierung aller bzw. ausschließlich nicht-synonymer Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,01 bzw. $\leq$ 0,05. Betrachtet wird der quantitative Phänotyp mit Kovariablen. Der Fehler 1. Art, die empirische und die minimale Teststärke wurden über 200 Replikate gemittelt. Bei einem erhöhten Fehler 1. Art sind die Werte der Teststärke als nicht-valide in Klammern angegeben.

		nicht-synonyme Varianten			alle Varianten		
MAF	Methode	Fehler 1.Art	Empirische Teststärke	Minimale Teststärke	Fehler 1.Art	Empirische Teststärke	Minimale Teststärke
$\leq$ 0,01	RVT1	0,09	(0,20)	(1,00)	0,10	(0,17)	(1,00)
	SKAT	0,09	(0,18)	(1,00)	0,10	(0,17)	(1,00)
	SKAT-O	0,10	(0,21)	(1,00)	0,10	(0,19)	(1,00)
$\leq$ 0,05	RVT1	0,10	(0,19)	(1,00)	0,11	(0,17)	(1,00)
	SKAT	0,09	(0,18)	(1,00)	0,10	(0,17)	(1,00)
	SKAT-O	0,10	(0,21)	(1,00)	0,11	(0,18)	(1,00)

RVT1: rare variant test 1; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT.

Die Quantil-Quantil-Plots in Abbildung 3.11 zeigen, dass für alle drei Gruppierungsmethoden die beobachteten und die erwarteten p-Werte stark voneinander abweichen, unabhängig von der Art der Gruppierung der ROI. Insbesondere werden die p-Werte im Median über die 200 Replikate stets überschätzt. Die Schwankung der p-Werte ist für alle hier betrachteten Gruppierungsmethoden von ähnlicher Größe, unabhängig von der Art der Gruppierung der ROI, steigt jedoch entgegengesetzt proportional zur Größe der p-Werte.

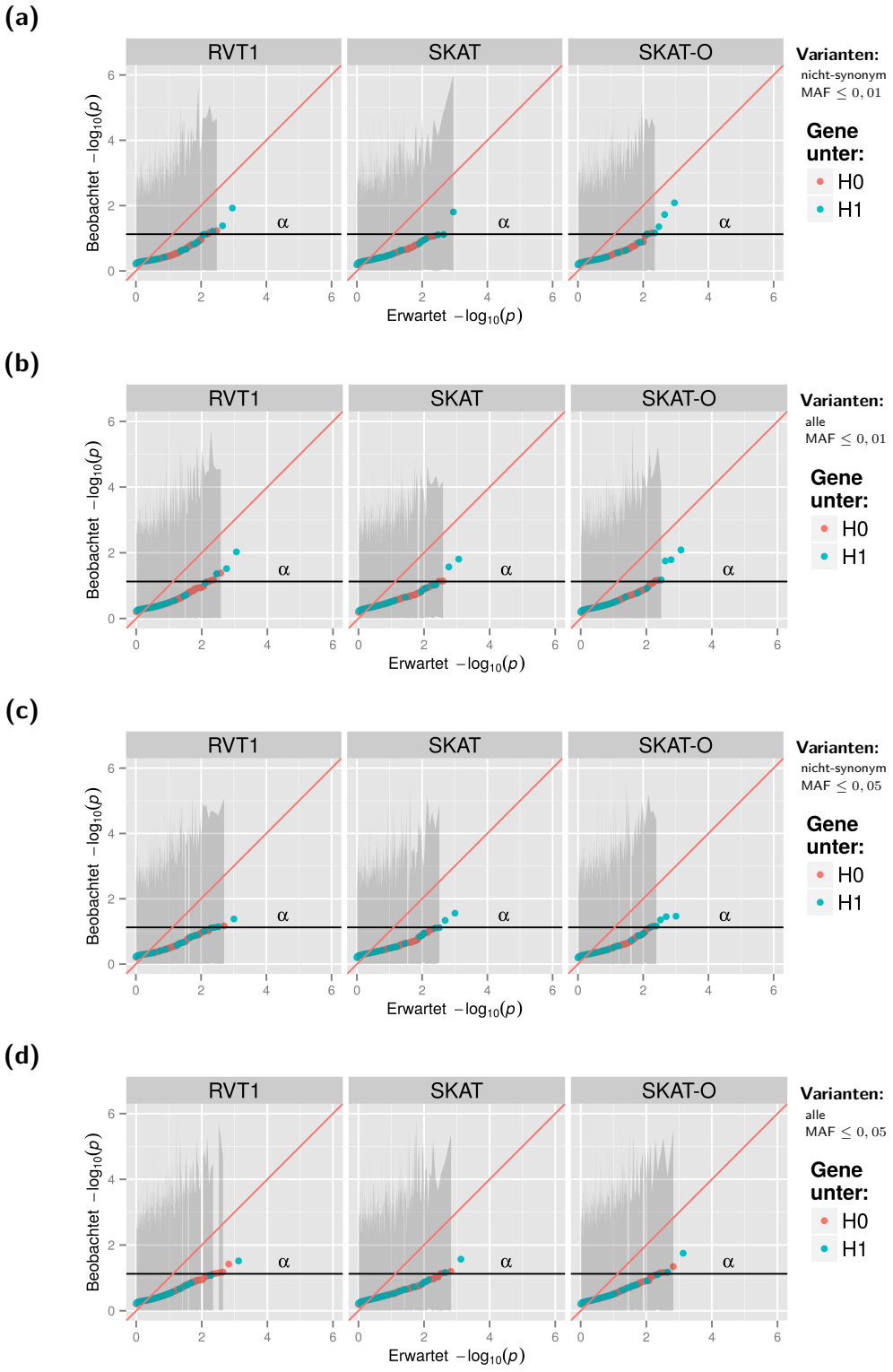


Abbildung 3.11: Quantil-Quantil-Plots der medianen zur Basis 10 log-normierten p-Werte gegen die zur Basis 10 log-normierten p-Werte beobachteten p-Werte innerhalb des ersten und dritten Quartils über 200 Replikate. Betrachtung des quantitativen Phänotyps mit Kovariablen bei Gen-weiser Gruppierung (a) ausschließlich nicht-synonymer und (b) aller Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,01 und Gen-weise Gruppierung (c) ausschließlich nicht-synonymer und (d) aller Varianten mit einer MAF $\leq$ 0,05.

RVT1: rare variant test 1; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT.

Die Streudiagramme in Abbildung 3.12 zeigen ähnliche Zusammenhänge zwischen den Variablen, Anteil der verursachenden Varianten, kumulierte MAF und kumuliertem Effekt bzgl. der empirischen Teststärke auf, wie bei der Untersuchung des quantitativen Phänotyps ohne Kovariablen in Abbildung 3.10.

3.7 Interpretation

Bei der Betrachtung des Fall-Kontroll-Status wies der Großteil der betrachteten Gruppierungsmethoden in den vier untersuchten Filter-Szenarien einen stark erhöhten Fehler 1. Art auf. Nur vier der betrachteten Methoden, C-α, FPCA, KBAC und VT hielten den Fehler 1. Art in allen untersuchten Filter-Szenarien ein und lieferten somit gültige Ergebnisse für die analysierten Teststärken. In keinem der analysierten Szenarien war die empirische Teststärke über 200 Replikate bei den validen Methoden größer als 0,13. Dabei lag die minimale Teststärke, also der Anteil der Replikate, in denen mindestens eine Assoziation erkannt wurde, immer zwischen 0,57 und 1,00. Unter den gültigen Methoden wies C-α in allen vier Gruppierungsszenarien sowohl die größte empirische wie auch minimale Teststärke auf. Der VT-Ansatz hatte zwar in allen Gruppierungsszenarien einen gültigen Fehler 1. Art, zeigte aber gleichzeitig vernachlässigbar geringe Werte für die empirische Teststärke, die nie größer als 0,04 und stets kleiner als der entsprechende Fehler 1. Art waren.

Weiterhin zeigte sich die Tendenz, dass eine Gruppierung nicht-synonymer Varianten bessere Ergebnisse bzgl. der Teststärke liefert, als eine Gruppierung aller Gen-Varianten. Dies entspricht der Erwartung, da bei der Erweiterung der ROIs von ausschließlich nicht-synonymen Varianten auf Varianten des gesamten Gens lediglich neutrale Varianten hinzukommen. Dies zeigt insbesondere, dass die Methoden nicht robust gegenüber dem Vorhandensein neutraler Varianten sind und somit eine Eingrenzung der relevanten Varianten bei der Bildung der ROIs unabdingbar ist. Ferner ergibt eine Gruppierung der Varianten mit MAF $\leq$ 0,05 eine höhere Teststärke als eine Gruppierung der Varianten mit MAF $\leq$ 0,01, was wiederum die These von Derkach et al. (2014) untermauert, dass die Teststärke neben anderen Faktoren von der MAF der Varianten abhängt.

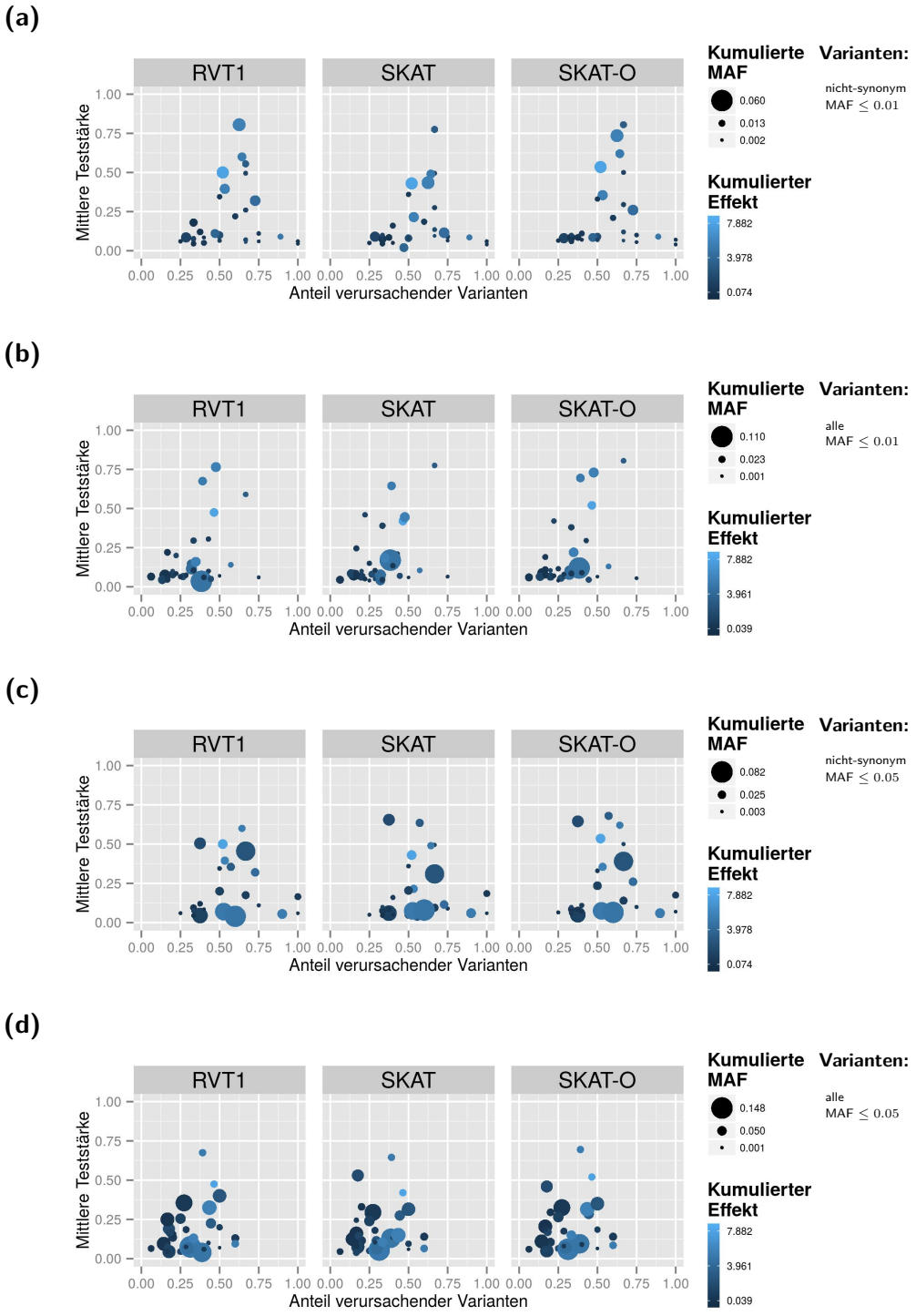


Abbildung 3.12: Streudiagramme der assoziierten ROIs bzgl. des Anteils der verursachenden Varianten und der empirischen Teststärke über 200 Replikate. Die Größe der Punkte ist proportional zur Summe der MAFs, je heller die Punkte, desto größer ist der aufsummierte Effekt der Varianten der jeweiligen ROI. Betrachtung des quantitativen Phänotyps mit Kovariablen bei Gen-weiser Gruppierung (a) ausschließlich nicht-synonymer und (b) aller Varianten mit einer Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,01 und Gen-weiser Gruppierung (c) ausschließlich nicht-synonymer und (d) aller Varianten mit einer MAF $\leq$ 0,05.
RVT1: rare variant test 1; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT.

Bei der Assoziationsuntersuchung des Fall-Kontroll-Status unter Berücksichtigung von Kovariablen zeigten SKAT und SKAT-O in allen Gruppierungsszenarien ungültige Ergebnisse, da der Fehler 1. Art stets zu hoch war. Außer bei der Gruppierung aller Gen-Varianten mit einer MAF $\leq$ 0,05 wies RVT1 in Gruppierungsszenarien einen gültigen Fehler 1. Art auf. Auch unter Verwendung von Kovariablen konnte keine der entsprechenden Gruppierungsmethoden überzeugen, weder für den binären noch für den quantitativen Phänotyp. Lediglich für die Betrachtung des binären Phänotyps mit Kovariablen lieferte die Methode RVT1 einen gültigen Fehler 1. Art in fast allen Filterszenarien. Bei den gültigen Filterszenarien hatte RVT1 im Vergleich zur Untersuchung des binären Phänotyps ohne Kovariablen eine höhere minimale Teststärke, jedoch war die empirische Teststärke etwas niedriger.

Bei der Assoziationsuntersuchung des quantitativen Phänotyps wiesen bis auf den VT-Ansatz alle betrachteten Gruppierungsmethoden einen zu hohen Fehler 1. Art auf; die Ergebnisse des VT-Ansatzes waren zwar gültig, die Werte der empirischen Teststärke mit niedrigeren Werten im Vergleich zum Fehler 1. Art jedoch vernachlässigbar klein. Sowohl der Wert der empirischen (0,04) als auch der minimalen Teststärke (0,85) waren für das Gruppierungsszenario mit allen Gen-Varianten und einer MAF $\leq$ 0,01 am größten für den VT-Ansatz.

Bei der Assoziationsuntersuchung des quantitativen Phänotyps unter Berücksichtigung von Kovariablen konnte keine der betrachteten Gruppierungsmethoden gültige Ergebnisse aufweisen, da der Fehler 1. Art stets zu hoch war.

Insgesamt zeigt sich bezogen auf den Simulationsdatensatz, dass eine gute Filterung der vorhandenen Informationen bzgl. der zur Verfügung stehenden Varianten nötig ist, um verlässliche Aussagen hinsichtlich einer untersuchten Assoziation zu einem gegebenen Phänotyp treffen zu können. Jede weitere Information, wie eine neutrale Variante oder eine Einflussvariable, birgt die Gefahr eines zusätzlichen „Rauschens“, was das Aufdecken einer vorhandenen Assoziation erschwert. So steigen zwar die betrachteten Teststärken, gleichzeitig aber auch der Wert des Fehlers 1. Art proportional zum Zuwachs an Informationen, d.h. mit der Höhe der gewählten MAF und dem Anteil neutraler Varianten innerhalb der ROIs. Auch verbessern sich die Testergebnisse nicht zwangsläufig, sobald Kovariablen in die Untersuchung einbezogen werden.

Für den betrachteten Datensatz ist keine klare Überlegenheit der auf linearen oder der quadratischen Teststatistiken basierten Gruppierungsmethoden zu erkennen. Die meisten gültigen Methoden verwenden Gewichte in der Teststatistik, was deren positiven Einfluss zeigt. Das Verhältnis zwischen den Permutations-basierten und den asymptotischen Ansätzen unter den gültigen Methoden ist ausgewogen.

Die Verteilung der p-Werte bzgl. der 200 Replikate entsprach bei keiner Methode für irgendein Szenario der Erwartung. Insbesondere die Methode PWST zeigte viel zu liberale p-Werte, d.h. unterschätzte die wahren p-Werte, über alle Replikate und Gruppierungsszenarien. Generell zeigten die in allen Gruppierungsszenarien gültigen Methoden die geringste Schwankung bzgl. des ersten und dritten Quartils der p-Werte über 200 Replikate. Unter den gültigen Methoden zeigt C-α die geringste Schwankung.

In allen untersuchten Szenarien ist das Zusammenspiel der drei Einflussfaktoren kumulierte MAF und kumulierter Effekt und Anteil der verursachenden Varianten auf die Größe der empirischen Teststärke erkennbar. Dabei bewirkt einzig die simultane Erhöhung aller drei Faktoren höhere Werte in der Teststärke. Auch dies stärkt die These, dass eine möglichst genaue Definition der geeigneten ROIs sowie eine moderate Stichprobengröße der Schlüssel für zuverlässige Aussagen über den Zusammenhang von seltenen Varianten und der Ausprägung eines bestimmten Phänotyps ist.

4 Anwendung: Das Gen *SLCO1B1* bei Leukämie

4.1 Einführung

Methotrexat ist ein Analogon zur Folsäure (Vitamin B9) und wird häufig als Zytostatikum in der Chemotherapie zur Behandlung von Krebserkrankungen eingesetzt. In einer genomweiten Assoziationsstudie konnten Treviño et al. (2009) einen signifikanten Zusammenhang zwischen einem reduzierten Methotrexat-Abbau bei Kindern mit akuter lymphatischer Leukämie (ALL) und häufigen Einzelbasenaustauschen im Gen *SLCO1B1* nachweisen. In einer sich anschließenden genomweiten Assoziationsstudie der gleichen Arbeitsgruppe konnten Ramsey et al. (2012) für eine erweiterte Stichprobe nachweisen, dass neben häufigen Varianten (MAF $> 0{,}05$) auch seltene Varianten des *SLCO1B1*-Gens einen Einfluss auf den Methotrexat-Abbau bei Kindern mit ALL haben. Dabei konnten Ramsey et al. (2012) insbesondere zeigen, dass die Effektstärken der seltenen Varianten wesentlich größer waren als die der häufigen Varianten.

Zur Qualitätssicherung sequenzierten Ramsey et al. (2012) die Genotypdaten des *SLCO1B1*-Gens mit verschiedenen Sequenzierungstechnologien. Für die Personen des ersten und letzten Dezils, d.h. für die oberen und unteren 10% an Personen mit niedrigem bzw. hohem Methotrexat-Abbau wurde zusätzlich eine Sanger-Sequenzierung, welche den Goldstandard in der DNA-Sequenzierung darstellt, durchgeführt (Ramsey et al. 2012). Anschließend wurden die Varianten hinsichtlich ihrer Funktionalität eingeordnet. Es ergab sich, dass von 93 Varianten 15 die zugehörige Aminosäure nicht-synonym kodieren. Um die Auswirkung der seltenen Varianten insbesondere

hinsichtlich ihrer Schädlichkeit vorherzusagen, wurden die Vorhersage-Algorithmen *SIFT* (Ng und Henikoff 2003), *PMUT* (Ferrer-Costa et al. 2004), *SNPS3D* (Yue et al. 2006) und *PolyPhen2* (Adzhubei et al. 2010) verwendet. Dabei wurde eine Variante als schädlich angesehen, wenn mindestens drei der vier vorgenannten Algorithmen diese Variante als schädlich deklariert hatten, was bei sieben der 15 nicht-synonymen Varianten der Fall war.

Auf Basis dieser Ergebnisse soll im Folgenden der Realdatensatz aus der Arbeit von Ramsey et al. (2012) mit den in dieser Arbeit betrachteten Gruppierungsmethoden untersucht werden. Hierfür standen die Genotyp-Daten von 93 Varianten von 673 Kindern verschiedenen Alters und verschiedenen Populationen zur Verfügung. Der quantitative pharmakogenetische Phänotyp des Methotrexat-Abbaus wurde durch Ramsey et al. (2012) für die Kovariablen Alter, Geschlecht, Populationszugehörigkeit sowie Behandlungsgruppe adjustiert. Für die Gruppierungsmethoden, die nur einen binären Phänotyp in der Untersuchung zulassen (vgl. Tabelle 2.1), wurden die Individuen des ersten und des letzten Dezils der adjustierten Methotrexat-Abbauerteilung verwendet, um zwei Gruppen zu erhalten. In diesem Fall reduzierte sich die Stichprobengröße auf 134 mit jeweils 67 Kindern je Gruppe. Für die Gruppierungsmethoden, die den p-Wert empirisch mittels Permutation schätzen, werden für jede der gebildeten Regionen von Interesse 10^9 Permutationen des Phänotyps durchgeführt.

4.2 Filterszenarien – Die Regionen von Interesse

Auf Basis der Genotypdaten des *SLCO1B1*-Gens aus der Arbeit von Ramsey et al. (2012) wurden in dieser Arbeit drei Regionen von Interesse gebildet: Varianten des gesamten Gens, ausschließlich nicht-synonyme Varianten des Gens und schädigende Varianten gemäß der Arbeit von Ramsey et al. (2012) zur genauen Definition der ROIs wurde nach betrachtetem Phänotyp unterschieden.

Für die Gruppierungsmethoden, die eine Untersuchung des quantitativen Phänotyps zulassen, wurden die Daten der gesamten Stichprobe verwendet, lediglich eine nicht-polymorphe Variante wurden entfernt, da in keiner der Probanden der Stichproben

mindestens ein seltenes Allel für diese Variante vorlag. Von den verbleibenden 92 Varianten waren 15 nicht-synonym und sieben schädigend. Schließlich ergab die Filterung nach den beiden MAF-Grenzen von 0,01 und 0,05 die ROIs, deren Anzahl in Tabelle 4.1 angegeben ist.

Varianten	# Schädigend	# Nicht-synonym	# Gesamtes Gen
alle	7	15	92
MAF $\leq$ 0,05	6	11	47
MAF $\leq$ 0,01	6	11	39

Tabelle 4.1: Anzahl (#) von Varianten in den drei Regionen von Interesse jeweils bestehend aus nur schädigenden, nicht-synonymen oder allen Gen-Varianten des Gens *SLCO1B1* bei Betrachtung des quantitativen Phänotyps, nach Entfernung nicht-polymorpher Varianten für die Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) $\leq$ 0,05 und $\leq$ 0,01.

Für die Gruppierungsmethoden, die nur die Betrachtung eines dichotomen Phänotyps zulassen, wurden zwei Gruppen hinsichtlich des ersten und letzten Dezils der adjustierten Methotrexat-Abbau-Verteilung gebildet. Dabei ergab sich, dass 19 der ursprünglichen 93 Varianten für diese Teilstichprobe als nicht-polymorph entfernt wurden, da sie in der verbliebenen Stichprobe von 134 Kindern nicht mehr vorkamen. Anschließend wurden ROIs bzgl. der MAF-Grenzen von 0,01 und 0,05 und den Untergruppen von schädigenden, nicht-synonymen und allen Varianten des *SLCO1B1*-Gens gebildet. Daraus ergaben sich die in Tabelle 4.2 angegeben Variantenzahlen.

Varianten	# Schädigend	# Nicht-synonym	# Gesamtes Gen
alle	6	12	74
MAF $\leq$ 0,05	5	9	29
MAF $\leq$ 0,01	5	8	20

Tabelle 4.2: Anzahl (#) von Varianten in den drei Regionen von Interesse bestehend aus nur schädigenden, nicht-synonymen oder allen Gen-Varianten des Gens *SLCO1B1* bei Betrachtung des dichotomen Phänotyps der Gruppen des ersten und letzten Dezils de adjustierten Methotrexat-Abbau-Verteilung, nach Entfernung nicht-polymorpher Varianten, für die Frequenz des seltenen Allels (engl. Minor Allele Frequency, MAF) von 0,05 und 0,01.

In den Tabellen 4.1 und 4.2 ist zu sehen, dass es nur geringe Abweichungen bei den Variantenzahlen in den ROIs bei Filterung nach den MAF-Grenzen von 0,01 und

0,05 gibt. In Anbetracht der langen Rechenzeiten bei den Permutationsansätzen werden deshalb nur die ROIs mit Varianten mit einer MAF $\leq 0{,}05$ untersucht.

4.3 Ergebnisse

In Tabelle 4.3 sind die p-Werte für die Gruppierungsmethoden für die drei ROIs des *SLCO1B1*-Gens bzgl. des Methotrexat-Abbaus bei Leukämie-kranken Kindern angegeben. Bei den Permutations-basierten Gruppierungsmethoden wurden für die Schätzung 10^9 Permutationen durchgeführt. Die ersten vier Zeilen der Tabelle geben die Ergebnisse der auf Basis der Simulationsdaten als gültig klassifizierten Methoden wieder.

Für die aus ausschließlich schädigenden Varianten bestehende ROI lieferten alle Gruppierungsmethoden außer CAST einen p-Wert kleiner als 0,05. Bei der Betrachtung von nicht-synonymen bzw. allen Gen-Varianten wiesen sechs bzw. vier Methoden einen p-Wert kleiner als 0,05 auf. Alle Methoden bis auf PWST, FPCA und SKAT hatten die kleinsten p-Werte für die ROI bestehend aus schädigenden Varianten. Die Mehrheit der Methoden hatte ihren jeweils höchsten p-Wert bei der Untersuchung nur nicht-synonymer Varianten. Außer C-α liefert keine gültige Methode einen signifikanten p-Wert bei der Gruppierung aller Gen-Varianten des Gens *SLCO1B1*. Der kleinste gültige p-Wert, wird bei der Gruppierung der schädigenden Varianten durch C-α mit $1{,}13 \cdot 10^{-21}$ erzielt, der nächst-kleinere p-Wert für ausschließlich schädigende Varianten von FPCA mit $6{,}75 \cdot 10^{-06}$. Für die Gruppierung ausschließlich nicht-synonymer Varianten des Gens *SLCO1B1* liefern C-α und FPCA mit $1{,}63 \cdot 10^{-05}$ bzw. $1{,}26 \cdot 10^{-06}$ signifikante p-Werte, während KBAC und der VT-Ansatz keine Assoziation erkennen. FPCA weist als einzige der vier gültigen Methoden bei der Gruppierung von nur nicht-synonymen Varianten des Gens *SLCO1B1* einen kleineren p-Wert als bei der Gruppierung von nur schädigenden Varianten auf. Generell gilt, dass die p-Werte für die Gruppierung ausschließlich schädigender Varianten des Gens *SLCO1B1* kleiner sind als für die Gruppierung nur nicht-synonymer und diese wiederum kleiner als für die Gruppierung aller Gen-Varianten sind.

Tabelle 4.3: p-Werte der 15 Gruppierungsmethoden aus der Assoziationsanalyse zwischen dem adjustierten Methotrexat-Abbau und den drei Regionen von Interesse von schädigenden, nicht-synonymen und allen Varianten mit einer MAF $\leq$ 0,05 des Gens *SLCO1B1* bei Leukämie-kranken Kindern.

Methode	schädigend	nicht-synonym	Gen-basiert
C-α	$1{,}13 \cdot 10^{-21}$	$1{,}63 \cdot 10^{-05}$	$9{,}93 \cdot 10^{-12}$
FPCA	$6{,}75 \cdot 10^{-06}$	$1{,}26 \cdot 10^{-06}$	$3{,}61 \cdot 10^{-01}$
KBAC	$2{,}89 \cdot 10^{-02}$	$4{,}46 \cdot 10^{-01}$	$3{,}44 \cdot 10^{-01}$
VT	$2{,}37 \cdot 10^{-04}$	$1{,}11 \cdot 10^{-01}$	$1{,}27 \cdot 10^{-01}$
ASUM	$5{,}30 \cdot 10^{-08}$	$1{,}14 \cdot 10^{-01}$	$1{,}02 \cdot 10^{-02}$
CAST	$5{,}79 \cdot 10^{-02}$	$7{,}91 \cdot 10^{-01}$	$6{,}02 \cdot 10^{-01}$
CMAT	$< 1{,}0 \cdot 10^{-09}$	$8{,}74 \cdot 10^{-01}$	$6{,}69 \cdot 10^{-01}$
CMC	$1{,}76 \cdot 10^{-07}$	$3{,}77 \cdot 10^{-06}$	$1{,}03 \cdot 10^{-02}$
PWST	$1{,}36 \cdot 10^{-03}$	$1{,}18 \cdot 10^{-04}$	$1{,}95 \cdot 10^{-07}$
RC	$6{,}30 \cdot 10^{-07}$	$7{,}84 \cdot 10^{-06}$	$2{,}91 \cdot 10^{-06}$
RVT1	$1{,}30 \cdot 10^{-03}$	$7{,}79 \cdot 10^{-01}$	$5{,}36 \cdot 10^{-01}$
RVT2	$2{,}00 \cdot 10^{-03}$	$3{,}80 \cdot 10^{-01}$	$3{,}71 \cdot 10^{-01}$
SKAT	$2{,}92 \cdot 10^{-02}$	$1{,}30 \cdot 10^{-02}$	$2{,}36 \cdot 10^{-01}$
SKAT-O	$2{,}38 \cdot 10^{-03}$	$5{,}77 \cdot 10^{-02}$	$4{,}04 \cdot 10^{-01}$
WSS	$< 1{,}0 \cdot 10^{-09}$	$6{,}13 \cdot 10^{-01}$	$1{,}97 \cdot 10^{-01}$

C-α: C-alpha-based Test; FPCA: functional principal component analysis; KBAC: kernel-based adaptive cluster; VT: variable threshold; ASUM: adaptive summation; CAST: cohort allelic sum test; CMAT: cumulative minor-allele test; CMC: combined multivariate cluster; PWST: p-value weighted sum test; RC: Rarecover; RVT: rare variant test 1 und 2; SKAT: sequencing kernel association test; SKAT-O: optimal unified SKAT; WSS: weighted sum statistic.

4.4 Interpretation

Bei der Analyse der drei ROIs für das Gen *SLCO1B1* zeigt sich, dass kaum eine der Gruppierungsmethoden in der Lage ist, eine Assoziation zwischen der Gruppe aller Gen-Varianten und dem Methotrexat-Abbau bei Patienten mit ALL zu erkennen. Von den gültigen Methoden zeigt hier lediglich C-α, wie auch für die übrigen beiden Regionen von Interesse, ein signifikantes Ergebnis. Dies zeigt deutlich, dass das Vorhandensein von neutralen Varianten in der Region von Interesse für fast alle Methoden ein immenses Problem darstellt. Anders als zu erwarten, sind die Gruppierungsmethoden auch bei der Betrachtung nur nicht-synonymer Varianten kaum in der

Lage, die bestehende Assoziation zu erkennen. Dies kann daran liegen, dass unter den nicht-synonymen auch Varianten mit vergleichsweise geringen bis moderaten Effekten zu finden sind, die den Effekt der schädlichen Varianten im Mittel deutlich verringern, (für die geschätzten Effektstärken vgl. Anhang von Ramsey et al. (2012)). Lediglich für die Region von Interesse, in der ausschließlich die als schädigend klassifizierten Varianten enthalten sind, liefern alle gültigen Methoden ein signifikantes Ergebnis. Wie dem Anhang der Arbeit von Ramsey et al. (2012) zu entnehmen ist, weisen alle hier beteiligten Varianten eine Effektstärke von höchstens -11 auf, was auf einen starken einseitigen Effekt schließen lässt. Dies erklärt insbesondere die kleineren p-Werte bei den linearen Teststatistiken bzw. den Burden-Tests gegenüber den quadratischen Tests, die insbesondere für die Untersuchung von ROIs mit bi-direktionale Effekten konstruiert wurden.

Insgesamt weisen die Gruppierungsmethoden, in denen der dichotomisierte Phänotyp (die beiden Gruppen bzgl. des unteren und oberen Dezils der adjustierten Methotrexat-Abbau-Verteilung) untersucht wurde, deutlich kleinere p-Werte auf als die Methoden, die den quantitativen Phänotyp untersuchen. Dies daran liegen, dass bei der Betrachtung des dichotomen Phänotyps nur die Individuen aus den Extremen der Verteilung des Methotrexat-Abbaus betrachtet werden. Die als schädigend eingestuften Varianten kommen im Wesentlichen bei den Individuen vor, deren Methotrexat-Abbaus gering ist. In dem Fall ist der Anteil der Individuen, für die möglicherweise kausale Varianten in den ROIs vorhanden sind, im Vergleich zur Gesamtstichprobe deutlich erhöht und somit ein Effekt möglicherweise besser detektierbar. Dies mag auch eine Begründung dafür sein, dass in den Filter-Szenarien der ROIs mit nicht-synonymen bzw. allen Gen-Varianten die Methoden, die den quantitativen Phänotyp verwenden, im Vergleich zu den Methoden, die den dichotomisierten Phänotyp untersuchen, deutlich höhere p-Werte aufweisen.

5 Diskussion und Ausblick

Gruppierungsmethoden erlauben die Untersuchung der Frage, ob ein Zusammenhang zwischen einer Gruppe von seltenen Varianten einer sog. Region von Interesse (ROI) und einer häufig vorkommenden Krankheit besteht. Für die Bildung der ROI wurde in dieser Arbeit erstmals eine detaillierte Anleitung gegeben und es wurden mögliche Kriterien sowie verschiedene Annotationwerkzeuge und -software diskutiert. Des weiteren wurden 15 ausgewählte Gruppierungsmethoden sowohl hinsichtlich ihrer definierenden Eigenschaften als auch im Hinblick auf ihre statistische Leistung auf Basis eines Simulationsdatensatzes miteinander verglichen. Die daraus gewonnenen Ergebnisse wurden anhand eines Realdatensatzes überprüft.

Bei der Untersuchung des Simulationsdatensatzes wurden zum Leistungsvergleich systematisch verschiedene Filterkriterien zur Bildung einer ROI verwendet. Die verschiedenen ROIs wurden zum einen über zwei MAF-Grenzen, zum anderen hinsichtlich der Auswirkung auf die Kodierung der entsprechenden Aminosäuren der Varianten eines Gens gebildet. Einige Gruppierungsmethoden erlauben eine Untersuchung von binären und quantitativen Phänotypen sowie die Einbeziehung vermuteter Einflussvariablen. Für die entsprechenden Methoden wurden diese Möglichkeiten zusätzlich berücksichtigt. Dies ist durchaus von Bedeutung, da der Fehler 1. Art und die Teststärke einer Methode von der Skalierung des betrachteten Phänotyps sowie der Berücksichtigung von vorhandenen Kovariablen abhängen können.

Ein Gesamtvergleich aller Gruppierungsmethoden hinsichtlich des Simulationsdatensatzes für beide Phänotypen sowohl mit als auch ohne Kovariablen ist nicht möglich, da nur die Methoden RVT1, SKAT und SKAT-O eine Untersuchung aller Szenarien zulassen und zugleich ein geeignetes Software-Programm für die Analyse vorhanden war. Es fällt auf, dass keine der drei Gruppierungsmethoden in jeder Situation gültig war. Die Methoden SKAT und SKAT-O weisen sogar in allen Untersuchungs-

und Filterszenarien einen erhöhten Fehler 1. Art auf, obwohl für diese Methoden eine Adjustierung wegen der geringen Stichprobengröße durchgeführt wurde, wie von Lee, Emond et al. (2012) und Wu et al. (2011a) empfohlen. Für die Methode RVT1 zeigen sich auf Ebene der Filterszenarien nahezu konstante Werte für den Fehler 1. Art bei der Betrachtung des binären Phänotyps, sowohl mit als auch ohne Kovariablen. Außerdem fällt auf, dass RVT1 bei der Verwendung von Kovariablen für jedes Filterszenario, in der verwendeten Reihenfolge, eine Steigerung der minimalen Teststärke erfährt, während die mittlere Teststärke jeweils leicht sinkt. Bei der analogen Betrachtung des quantitativen Phänotyps mit und ohne Kovariablen für die Methode RVT1 sieht man, dass sowohl der Fehler 1. Art als auch die Werte der Teststärken stets konstant bleiben. Allerdings ist zu beachten, dass der Fehler 1. Art bei der Betrachtung des quantitativen Phänotyps für RVT1 in allen Filterszenarien stets erhöht ist, so dass aus dieser Betrachtung keine verlässlichen Schlussfolgerungen gezogen werden können.

Ein Leistungsvergleich unter annähernd gleichen Bedingungen unter Verwendung des Simulationsdatensatzes ist zumindest für die meisten der 15 Gruppierungsmethoden bei Betrachtung des Fall-Kontroll-Status ohne Kovariablen möglich. Die Methoden VT und PWST lassen allerdings auf Grund der langen Rechenzeit nur die Untersuchung des quantitativen Phänotyps zu. In diesem Gesamtvergleich gibt es nur vier Methoden, für die alle vier Filterszenarien gültig sind: C-α, FPCA, KBAC und VT, wobei C-α stets die größte mittlere und minimale Teststärke aufweist. VT ist zwar auch stets gültig, hat aber vernachlässigbar kleine Werte für die mittlere Teststärke, die nie größer als 0,04 und gleichzeitig immer kleiner als der entsprechende Fehler 1. Art ist.

Einige Methoden fallen besonders auf, da sie in allen Szenarien, in denen sie angewendet werden können, einen deutlich erhöhten Fehler 1. Art aufweisen: SKAT, SKAT-O, CMC, ASUM und PWST. Unter diesen ungültigen Methoden sticht PWST mit einem Fehler 1. Art zwischen 0,50 und 0,65 deutlich hervor. Auch eine höhere Stichproben- oder Permutationszahl in der Schätzung der jeweiligen p-Werte würde diesen Effekt vermutlich nicht beeinflussen, da die Betrachtung der p-Wert-Verteilung von PWST bzgl. der Simulationsstudie eine ganz andere Form zeigt als erwartet.

Insgesamt zeigt sich aus der Verwendung des Simulationsdatensatzes, dass die Teststärken der meisten Gruppierungsmethoden maßgeblich vom Anteil der verursachenden Varianten, dem kumulierten genetischen Effekt und der kumulierten Frequenz des seltenen Allels abhängen. Dies bestätigt die Thesen der Arbeiten von Derkach et al. (2014) und Ladouceur et al. (2012). Des Weiteren bestärkt dies sowohl die Notwendigkeit einer ausführlichen Literaturrecherche und die Verwendung von vorhandenem Vorwissen als auch die Annotation der zur Verfügung stehenden Genotypdaten der seltenen Varianten mit relevanten Informationen, um relevante Filterkriterien zu finden und anwenden zu können.

Die Untersuchung des Realdatensatzes zeigt, dass bis auf CAST alle Gruppierungsmethoden in der Lage sind, eine Assoziation zu erkennen, falls eine geeignete Filterung erfolgt. Hierbei erweisen sich die Filterungen nach der Frequenz des seltenen Allels und nach der Klassifizierung hinsichtlich der Schädlichkeit der Varianten als besonders wichtig.

Insgesamt zeigen die Ergebnisse der Simulationsstudie, dass keine der betrachteten Gruppierungsmethoden für alle Situationen empfehlenswert ist, da sie entweder den Fehler 1. Art nicht einhalten oder eine zu geringe Teststärke aufweisen. Die Gültigkeit der Methoden hängt dabei z.T. davon ab, wie der betrachtete Phänotyp skaliert ist oder ob Kovariablen einbezogen werden.

Eine weitere wichtige Folgerung aus den vorliegenden Ergebnissen ist, dass die Detektion einer vorhandenen Assoziation durch die zusätzliche Präsenz von neutralen Varianten erschwert wird. Dies gilt sogar dann, wenn die Effektstärken der kausalen Varianten sehr groß sind. Aus diesem Grund stellt die Annotation und die Filterung nach geeigneten Kriterien nicht nur die Basis für die Bildung einer geeigneten Region von Interesse dar, sondern ist auch essentiell für die Durchführung einer aussagekräftigen Assoziationsstudie.

Bei geringem oder fehlendem Vorwissen erweist sich eine Filterung nach der Frequenz des seltenen Allels und nach der Funktionalität in Form einer Annotation über die Kodierung der entsprechenden Aminosäuren der Varianten als geeignete Kriterien, um den Großteil nicht-relevanter Varianten zu entfernen. Eine zusätzliche Annotation oder Schätzung von Maßen zur Bewertung der Schädlichkeit der Varianten einer ROI

kann hilfreich sein, um die Teststärke zu erhöhen. Dabei muss aber beachtet werden, dass Annotationen aus bestehenden Datenbanken in Situationen wie dem alternativen Splicen zu uneindeutigen Ergebnissen führen können. So geben Stitziel et al. (2011) zu bedenken, dass eine Annotation fehlerhaft und insbesondere nicht eindeutig sein kann, falls eine Variante zu mehreren Transkripten eines Gens gehört und sie somit unter Umständen verschieden auf die entsprechenden Aminosäuren wirkt. Daher ist die ständige Aktualisierung der Annotationsdatenbanken insbesondere im Hinblick auf die Informationen über seltene Varianten eine große Herausforderung.

Für die Untersuchung seltener Varianten mit Hilfe von Gruppierungsmethoden lässt sich ein sinnvolles Vorgehen nicht allgemeingültig bestimmen. Das Vorgehen hängt von vielerlei Faktoren ab. Zum einen besteht eine starke Abhängigkeit von der gegebenen Datenbasis der Stichprobe: Welcher Phänotyp wird als assoziiert angenommen? Wie ist dieser skaliert? Existieren zusätzliche Informationen über mögliche Einflussfaktoren? Auf Basis dieser Gegebenheiten sollten geeignete Gruppierungsmethoden ausgewählt werden, für die möglichst auch eine validierte Software verfügbar ist. Liegen keine Informationen bzw. Indizien über die genetische Architektur der möglichen Assoziation vor, sollten sowohl auf linearen als auch auf quadratischen Teststatistiken basierende Gruppierungsmethoden angewendet werden da so ein breiteres Spektrum von Alternativhypothesen abgedeckt ist. Auch die Einbindung von häufigen Varianten in die Region von Interesse über eine geeignete Gruppierungsmethode oder eine Kombinationsregel wie die Fisher-Regel sollten in Erwägung gezogen werden, falls keine Informationen über die genetische Basis der Assoziation vorliegen. Die Anwendung sowohl von asymptotischen als auch von Permutations-basierten Tests wurde in der Literatur weithin diskutiert (vgl. Abschnitt 2.2.6). Es zeigt sich, dass eine bedenkenlose Anwendung nur eines einzelnen Ansatzes nicht zu empfehlen ist. Da zu jedem asymptotischen Test immer die Schätzung der p-Werte über Permutationen möglich ist, sollte eine solche Schätzung auch durchgeführt werden, um die Konkordanz der Ergebnisse beider Ansätze überprüfen zu können. Dabei bleibt zu beachten, dass die Schätzung exakter p-Werte sehr rechenintensiv sein kann, insbesondere bei genomweiten Studien.

Aufgrund der wenig zufriedenstellenden Ergebnissen dieser Studie im Hinblick auf die Einsetzbarkeit der Gruppierungsmethoden stellt sich die Frage nach geeigneten Verbesserungen der Ansätze oder sogar nach alternativen Methoden zur Untersuchung

des Zusammenhangs von häufigen Krankheiten und seltenen Varianten. Burkett und Greenwood (2013) schlagen vor, etablierte Methoden für die Untersuchung der Assoziation häufiger Varianten mit häufigen Krankheiten auch für seltene Varianten in Betracht zu ziehen. In den Arbeiten von Kinnamon et al. (2012) und Xu et al. (2012) konnte die Idee dieses Vorgehens für den Cochran–Armitage Test bzw. für einige Regressions-basierte Tests bestätigt werden.

Die Anzahl von Gruppierungsmethoden wächst ständig weiter, vor allem im Bereich der quadratischen Teststatistiken. Dabei werden oft Erweiterungen der SKAT-Ansätze von Lee, Emond et al. (2012) und Wu et al. (2011a) vorgeschlagen. Eine der jüngsten in diesem Zusammenhang beschriebenen Methoden stammen von Wei et al. (2014) und Wu et al. (2015). Ein Leistungsvergleich dieser neuen Methoden scheint sinnvoll und notwendig, insbesondere mit Blick auf die unbefriedigende Leistung von SKAT und SKAT-O im Rahmen dieser Arbeit.

Auch der Ansatz der Meta-Analyse, der in der Vergangenheit vor allem bei der Assoziationsuntersuchung von SNPs mit häufigen komplexen Krankheiten angewendet wurde, rückt für seltene Varianten in den Fokus. Aufgrund der nach wie vor sinkenden Preise in der Genom- und Exom-Sequenzierung ist es in den vergangen Jahren möglich geworden, immer größere Stichprobenumfänge in Assoziationsstudien zu untersuchen, und die Zahl der Veröffentlichungen von Sequenzierungs-Studien mit seltenen Varianten steigt kontinuierlich. Da für die Detektion von vorhandenen Assoziationen insbesondere im Zusammenhang mit seltenen Varianten große Stichprobenumfänge von Vorteil sind, ist die Entwicklung entsprechender Meta-Analyse-Methoden naheliegend. Einer der ersten in diesem Zusammenhang entstanden Ansätze wurde von Lumley et al. (2012) vorgeschlagen. Er basiert darauf, dass aus jeder beteiligten Studie Parameter aus der jeweiligen SKAT-Teststatistik von Wu et al. (2011a) eingebunden und zu einer neuen gemeinsamen Teststatistik kombiniert werden. Lee et al. (2013) schlagen eine MetaSKAT genannte Methode vor, in der mehrere Studien über ROI-basierte Teststatistiken für seltene Varianten miteinander kombiniert werden. Sie zeigen in ihrer Arbeit, dass dieser Ansatz ebenso gute Ergebnisse liefert, wie eine direkte Auswertung einer Studie mit der gleichen Stichprobengröße. In der jüngst erschienen Arbeit von Tang und Lin (2015) werden dieser und weitere Meta-analytische Ansätze (Feng et al. 2014; Lee et al. 2013; Tang und Lin 2013) für seltene Varianten miteinander verglichen. Außerdem entwickelten Tang und Lin

(2015) ein Software-Programm, das die Verwendung der Summary-Statistiken aller vier genannten Methoden in den Meta-Analysen der jeweils anderen Methode ermöglicht. Eine solche Software stellt vor allem für weltweit operierende Konsortien eine große Bereicherung und Flexibilität in der Bereitstellung der notwendigen Daten, dar.

Parallel zur Entwicklung der Gruppierungsmethoden sind in jüngster Zeit eine Reihe weiterer alternativer Ansätze zur Untersuchung der RVCD-Theorie veröffentlicht worden, darunter auch Familien-basierte Verfahren, bei denen im Gegensatz zu den Gruppierungsmethoden seltene Varianten als mögliche Ursache für eine Krankheit in Familien anstelle von unabhängigen Stichproben untersucht werden. Erste Methoden dieser Art wurden von Chen et al. (2013), De et al. (2013), Schifano et al. (2012) und Shugart et al. (2012) vorgeschlagen. Diese Ansätze sind aus mindestens zwei Gründen sinnvoll: Zum einen ist eine Häufung Krankheits-verursachender Varianten in Familien wesentlich wahrscheinlicher als in Gruppen nicht verwandter Personen. Zum anderen sind diese Ansätze weniger anfällig für Probleme die im Zusammenhang mit Populations-Stratifikation entstehen können. **feng__methods__2015** stellen in ihrem jüngst erschienen Artikel eine Erweiterung des SKAT-Ansatzes von Wu et al. (2011a) für die Untersuchung Familien-basierter Daten vor. Dabei kombinieren sie ihren Ansatz mit einer Reihe von den in der vorliegenden Arbeit untersuchten Gruppierungsansätzen und zeigen, dass die Teststärke ihres Ansatzes in bestimmten Situationen deutlich bessere Werte aufweist als in Studien mit nicht-verwandten Personen. Außerdem zeigen **feng__methods__2015**, dass ihre Methode leicht in den Kontext einer Meta-Analyse eingebettet werden kann. **epstein__statistical__2015** schlagen dazu einen ähnlichen Methoden-Komplex zur Untersuchung seltener Varianten in Familien vor, bei dem nur erkrankte Familienmitglieder in die Analyse miteingehen und keine Kontrollen benötigt werden. In einem weiteren Familien-basierten Ansatz von **lin__robust__2015** wird vorgeschlagen, chromosomale Regionen zwischen Geschwistern zu vergleichen und dabei sich unterscheidende Regionen als sogenannte „Fall"- und überstimmende als „Kontroll"-Regionen zu identifizieren.

In Haplotyp-basierten Ansätzen werden Haplotypen hinsichtlich der Assoziation mit einer häufigen Krankheit untersucht. Aus der Betrachtung einer Allel-Folge werden u.a. direkte Schlussfolgerungen über Variationen der entsprechenden Aminosäuren, der zugehörigen Gene und somit Proteinprodukten gezogen (Yang et al. 2008). Daher

ist es bei Haplotyp-basierten Ansätzen wichtig zu wissen, ob die Allele der betrachteten Varianten zusammen auf einem der beiden paarweisen Chromosomen vererbt wurden. Um das Potential der Daten aus bereits erfolgten GWAS auch für seltene Varianten auszuschöpfen, schlugen Li et al. (2010) zwei Ansätze zur Haplotyp-basierten Untersuchung von seltenen Varianten vor. Es zeigte sich, dass diese Ansätze auf der GWAS-Ebene bessere Ergebnisse erzielen konnten als die Gruppierungsmethoden WSS (Madsen und Browning 2009) und CMC von,(Li und Leal 2008). Dennoch gilt dies nicht uneingeschränkt für Sequenz-basierte Daten, wie Li et al. (2010) betonen. In einer 2015 erschienenen Arbeit vergleichen Wang und Lin (2015) die Gruppierungsmethoden CMC (Li und Leal 2008) und SKAT (Wu et al. 2011a) mit drei Haplotyp-basierten Ansätzen. Wang und Lin (2015) konnten zeigen, dass diese Ansätze insbesondere im Falle von Stichproben mit Populations-Stratifikation oder beim Vorhandensein von Varianten mit unterschiedlichen Effektrichtungen in der betrachteten Region den Fehler 1. Art besser einhalten als die in dieser Arbeit betrachteten Gruppierungsmethoden. Ein weiterer Vorteil der Haplotyp-basierten Ansätze ist, dass in Fall-Kontroll-Studien die Schätzung des Chancenverhältnisses zur erkranken für jeden einzelnen Haplotyp möglich ist, während selbst bei geeigneten Gruppierungsmethoden höchstens ein Effekt der Region geschätzt werden kann. Chen et al. (2015) schlugen erst vor kurzem eine Methode vor, die mit Genotyp-Daten sowohl in Phase, als auch in nicht-Phase und insbesondere mit Populations-stratifizierten Daten umgehen kann. In Phase heißt, dass für die betrachteten Genotypen jeweils bekannt ist, auf welchem der beiden Chromosomen die zugehörigen Allele vererbt wurden. Auch sind Ansätze zur Untersuchung von Haplotypen denkbar, die über die Gen-kodierenden Regionen hinausgehen, so dass auch Intron-Abschnitte und *cis/trans* Aktivitäten betrachtet werden könnten.

Bei den in dieser Arbeit betrachteten Gruppierungsansätzen lag der Fokus auf der Untersuchung ausschließlich autosomaler, also nicht-Geschlechts-spezifischer seltener Varianten. Neben dem Problem der unterschiedlichen Genotyp-Kodierung für Männer und Frauen bilden auch die Allelfrequenz-Schätzung, der Test auf das Hardy-Weinberg-Gleichgewicht und die Imputation fehlender Genotypen zusätzliche Schwierigkeiten gegenüber den nicht-Geschlechts-spezifischen Analysen. Um diese Lücke in der Analyse seltener Varianten auf den Geschlechts-Chromosomen zu schließen, stellten Ma et al. (2015) vor kurzem drei Gruppierungsmethoden für die Untersuchung seltener Varianten des X-Chromosoms vor. Diese Gruppierungs-

methoden sind jeweils an die Methoden WSS von Madsen und Browning (2009) sowie SKAT (Wu et al. 2011a) und SKAT-O (Lee, Emond et al. 2012) angelehnt. Ma et al. (2015) zeigen, dass ihre Ansätze für eine Reihe von Untersuchungsszenarien gute Teststärken aufweisen, weisen allerdings darauf hin, dass vor jeder Analyse seltener Varianten des X-Chromosoms zusätzliche Informationen notwendig sind um die richtige Genotyp-Kodierung in der Untersuchung wählen zu können, da die Teststärke ansonsten geringer ausfällt.

Die Weiterentwicklung der Sequenzierungstechniken ermöglicht mittlerweile die Sequenzierung vollständiger menschlicher Genome innerhalb von Tagen (Mardis 2013). Parallel dazu sinken auch die Kosten zur Sequenzierung eines Genoms ständig weiter. Die Kosten für die Genom-Sequenzierung einer Person lagen vor zwei Jahren noch bei ca. 5000$, mittlerweile sind es nur noch 1000-1500$ (Dijk et al. 2014; Wetterstrand 2015). Dies führt nicht nur zu einer großen Menge an neu generierten Daten, sondern auch zum Aufkommen einer Reihe von neuen medizinischen Fragestellungen sowie zu neuen potentiellen Behandlungsmöglichkeiten, nicht nur für bestimmte Gruppen von Patienten, sondern sogar für Einzelpersonen.

Zwar war die Entwicklung der Sequenzierungstechniken der nächsten Generation ein Durchbruch in der DNA-Sequenzierung. Allerdings sind die neuen Sequenzierungs-Technologien leider nicht fehlerfrei (Mardis 2013). Schon bei der Erstellung der Bibliotheken, in denen eine Vielzahl von Basensequenzen, die sog. Reads, erzeugt bzw. kopiert werden, treten Fehler auf. So können bestimmte Basensequenzen in der Amplifizierung gegenüber anderen durch die Polymerase bevorzugt werden oder es kann an einzelnen Positionen zum Einbau des falschen Nukleotids kommen. Diese Fehler werden durch den Prozess der Amplifizierung wiederum vervielfältigt (Mardis 2013). Dies führt bei einer folgenden Analyse evtl. zu einem erhöhten Fehler 1. Art. Die Einzel-Molekül-Sequenzierung, bei der die Sequenzierung einzelner DNA-Moleküle ohne Amplifizierung erfolgt, stellt eine mögliche Alternative zu den auf Amplifizierung basierenden Technologien dar. Jedoch sieht Mardis (2013) dabei noch die größte Herausforderung in der Weiterentwicklung der Detektoren, um die geringen Signalintensitäten eines einzelnen DNA-Moleküls messen zu können.

Die Weiterentwicklung der Sequenzierungstechniken und der Analysemethoden schafft gute Voraussetzungen für die immer bessere Untersuchung der Assoziation von sel-

tenen Varianten mit häufigen Krankheiten. Damit einher gehen aber auch neue Herausforderungen: die effiziente Verarbeitung der zunehmenden Datenmengen; Verbesserung, Sicherung und Kontrolle der Qualität der Sequenzierungsdaten und die Entwicklung und Leistungsbewertung neuer statistischer Analyseansätze. Unabhängige Vergleichsstudien sind für die Leistungsbewertung unverzichtbar. In diesem Kontext bietet die vorliegende Arbeit eine umfangreiche Studie zu aktuellen und häufig angewendeten Gruppierungsansätzen zur Untersuchung von Assoziationen zwischen seltenen Varianten und häufigen Erkrankungen. Es ist eine fortwährende Aufgabe, solche Vergleichsstudien auch für die neu hinzukommenden Methoden durchzuführen.

6 Zusammenfassung

In dieser Arbeit wurde eine systematische Analyse und ausführliche Leistungsbewertung von 15 Gruppierungsmethoden zur Untersuchung der Assoziationen zwischen Gruppen seltener Varianten und komplexen Krankheiten vorgenommen.

Obwohl in der Regel auf Basis eines Gens definiert, ist die genaue Zusammensetzung einer Region von Interesse a priori unbekannt. In dieser Arbeit wurde daher erstmals auf verschiedene Vorgehensweisen für ihre Bildung eingegangen und ein Schema zum Ablauf der Filterung angegeben. Gruppierungsmethoden werden zur Untersuchung eines möglichen Zusammenhangs zwischen einer Region von Interesse und der Ausprägung eines Phänotyps verwendet. Zur besseren Differenzierung der betrachteten Ansätze wurden allgemeine Charakteristika von Gruppierungsmethoden erläutert, die untersuchten Gruppierungsmethoden dahingehend eingeordnet und durch Algorithmen beschrieben. Ferner wurden unter Verwendung von Simulationsdaten und einem Realdatensatz die statistischen Eigenschaften Fehler 1. Art und Teststärke untersucht. Dabei wurden insbesondere die Verwendung von Kovariablen und die Methode zur Bildung der Region von Interesse variiert.

Keine der 15 Gruppierungsmethoden ist pauschal für alle Situationen empfehlenswert, da entweder der Fehler 1. Art nicht eingehalten wird oder eine zu geringe Teststärke vorliegt. Die Gültigkeit der Methoden hängt z.T. davon ab wie der betrachtete Phänotyp skaliert ist und ob Kovariablen einbezogen werden. Die Untersuchung des binären Phänotyps ohne Einbeziehung von Kovariablen erwies sich als das Szenario, in dem die meisten Methoden gültig sind, wobei mitunter keine ausreichend hohen Werte für die betrachteten Teststärken vorlagen.

Bei Betrachtung der verschiedenen Filterszenarien zeigt sich, dass die Zusammensetzung der Region von Interesse essentiell ist, um eine vorhandene Assoziation detektieren zu können. Bei geringem oder fehlendem Vorwissen erweisen sich eine Filterung nach der Frequenz des seltenen Allels und nach der Auswirkung auf die Kodierung der entsprechenden Aminosäuren der Varianten als geeignete Kriterien. Mit diesen Filtern kann ein Großteil der nicht-relevanten Varianten aus der Region von Interesse entfernt werden. Aus der Untersuchung des Realdatensatzes wurde zudem deutlich, dass eine zusätzliche Annotation von Maßen zur Bewertung der Schädlichkeit der einzelnen Varianten einer Region von Interesse hilfreich ist, um die Teststärken der meisten untersuchten Gruppierungsmethoden zu erhöhen.

Es ist zu vermuten, dass einige der wenig zufriedenstellenden Ergebnisse im Bereich der Leistungskriterien auf die vergleichsweise geringe Stichprobengröße zurückzuführen sind. Die zu geringen Teststärken haben zudem vermutlich ihren Grund in den kleinen Effektstärken der einzelnen Varianten im Zusammenspiel mit der wiederum geringen Stichprobengröße. Es ist anzunehmen, dass in naher Zukunft durch eine effizientere und kostengünstigere Genomsequenzierung die Untersuchung größerer Stichproben möglich wird und sich die Leistung der Gruppierungsmethoden weiter verbessert.

Literatur

1000 Genomes Project Consortium, Abecasis, GR, Altshuler, D, Auton, A, Brooks, LD, Durbin, RM, Gibbs, RA, Hurles, ME und McVean, GA (2010). „A map of human genome variation from population-scale sequencing“. *Nature* 467, S. 1061–1073.

Abramowitz, M und Stegun, IA (1965). „Handbook of mathematical functions, graphs, and mathematical tables“. *Appl Math Set* 55.

Adzhubei, IA, Schmidt, S, Peshkin, L, Ramensky, VE, Gerasimova, A, Bork, P, Kondrashov, AS und Sunyaev, SR (2010). „A method and server for predicting damaging missense mutations“. *Nat Methods* 7, S. 248–249.

Agresti, A (1992). „A survey of exact inference for contingency tables“. *Statistical Sci* 7, S. 131–153.

Ahituv, N, Kavaslar, N, Schackwitz, W, Ustaszewska, A, Martin, J, Hébert, S, Doelle, H, Ersoy, B, Kryukov, G, Schmidt, S, Yosef, N, Ruppin, E, Sharan, R, Vaisse, C, Sunyaev, S, Dent, R, Cohen, J, McPherson, R und Pennacchio, LA (2007). „Medical sequencing at the extremes of human body mass“. *Am J Hum Genet* 80, S. 779–791.

Almasy, L, Dyer, TD, Peralta, JM, Kent Jr, JW, Charlesworth, JC, Curran, JE und Blangero, J (2011). „Genetic Analysis Workshop 17 mini-exome simulation“. *BMC proceedings* 5 Suppl 9, S2.

Asimit, J und Zeggini, E (2009). „Testing for rare variant associations in complex diseases“. *Genome Med* 1, S. 24.

Auer, PL und Lettre, G (2015). „Rare variant association studies: considerations, challenges and opportunities“. *Genome Medicine* 7, S. 16.

Bacanu, SA, Nelson, MR und Whittaker, JC (2011). „Comparison of methods and sampling designs to test for association between rare variants and quantitative traits“. *Genet Epidemiol* 35, S. 226–235.

Bansal, V, Libiger, O, Torkamani, A und Schork, NJ (2010). „Statistical analysis strategies for association studies involving rare variants“. *Nat Rev Genet* 11, S. 773–785.

Barnett, IJ, Lee, S und Lin, X (2013). „Detecting rare variant effects using extreme phenotype sampling in sequencing association studies“. *Genet Epidemiol* 37, S. 142–151.

Basu, S und Pan, W (2011). „Comparison of statistical tests for disease association with rare variants“. *Genet Epidemiol* 35, S. 606–619.

Bennett, S (2004). „Solexa Ltd“. *Pharmacogenomics* 5, S. 433–438.

Bentley, DR et al. (2008). „Accurate whole human genome sequencing using reversible terminator chemistry“. *Nature* 456, S. 53–59.

Bera, AK und Bilias, Y (2001). „Rao's score, Neyman's C-(α) and Silvey's LM tests: An essay on historical developments and some new results“. *J Stat Plan Inference* 97, S. 9–44.

Bhatia, G, Bansal, V, Harismendy, O, Schork, NJ, Topol, EJ, Frazer, K und Bafna, V (2010). „A covering method for detecting genetic associations between rare variants and common phenotypes“. *PLoS Comput Biol* 6, e1000954.

Bickeböller, H und Thompson, EA (1996). „The probability distribution of the amount of an individual's genome surviving to the following generation“. *Genetics* 143, S. 1043–1049.

Bodmer, W und Bonilla, C (2008). „Common and rare variants in multifactorial susceptibility to common diseases“. *Nat Genet* 40, S. 695–701.

Bradley, JV (1978). „Robustness?“ *Br J Math Stat Psychol* 31, S. 144–152.

Brunham, LR, Singaraja, RR und Hayden, MR (2006). „Variations on a gene: Rare and common variants in ABCA1 and their impact on HDL cholesterol levels and atherosclerosis“. *Annu Rev Nutr* 26, S. 105–129.

Burkett, K und Greenwood, C (2013). „A sequence of methodological changes due to sequencing“. *Curr Opin Allergy Clin Immunol* 13, S. 470–477.

Byrnes, AE, Wu, MC, Wright, FA, Li, M und Li, Y (2013). „The value of statistical or bioinformatics annotation for rare variant association with quantitative trait“. *Genet Epidemiol* 37, S. 666–674.

Campbell, NA und Reece, JB (2000). *Biologie*. 2. Aufl. Heidelberg: Spektrum Akademischer Verlag. ISBN: 978-3-8274-0032-5.

Chen, H, Meigs, JB und Dupuis, J (2013). „Sequence kernel association test for quantitative traits in family samples“. *Genet Epidemiol* 37, S. 196–204.

Chen, R, Wei, Q, Zhan, X, Zhong, X, Sutcliffe, JS, Cox, NJ, Cook, EH, Li, C, Chen, W und Li, B (2015). „A haplotype-based framework for group-wise transmission/disequilibrium tests for rare variant association analysis“. *Bioinformatics* 31, S. 1452–1459.

Cingolani, P, Platts, A, Wang, LL, Coon, M, Nguyen, T, Wang, L, Land, SJ, Lu, X und Ruden, DM (2012). „A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3“. *Fly (Austin)* 6, S. 80–92.

Cohen, JC, Kiss, RS, Pertsemlidis, A, Marcel, YL, McPherson, R und Hobbs, HH (2004). „Multiple rare alleles contribute to low plasma levels of HDL cholesterol“. *Science* 305, S. 869–872.

Cordell, HJ (2009). „Detecting gene-gene interactions that underlie human diseases“. *Nat Rev Genet* 10, S. 392–404.

Cunningham, F, Amode, MR, Barrell, D, Beal, K, Billis, K, Brent, S, Carvalho-Silva, D, Clapham, P, Coates, G, Fitzgerald, S, Gil, L, Girón, CG, Gordon, L, Hourlier, T, Hunt, SE, Janacek, SH, Johnson, N, Juettemann, T, Kähäri, AK, Keenan, S, Martin, FJ, Maurel, T, McLaren, W, Murphy, DN, Nag, R, Overduin, B, Parker, A, Patricio, M, Perry, E, Pignatelli, M, Riat, HS, Sheppard, D, Taylor, K, Thormann, A, Vullo, A, Wilder, SP, Zadissa, A, Aken, BL, Birney, E, Harrow, J, Kinsella, R, Muffato, M, Ruffier, M, Searle, SMJ, Spudich, G, Trevanion, SJ, Yates, A, Zerbino, DR und Flicek, P (2015). „Ensembl 2015“. *Nucleic Acids Res* 43, S. D662–669.

De, G, Yip, WK, Ionita-Laza, I und Laird, N (2013). „Rare variant analysis for family-based design“. *PLoS One* 8, e48495.

Dering, C, Hemmelmann, C, Pugh, E und Ziegler, A (2011). „Statistical analysis of rare sequence variants: An overview of collapsing methods“. *Genet Epidemiol* 35 Suppl 1, S12–17.

Dering, C, König, IR, Ramsey, LB, Relling, MV, Yang, W und Ziegler, A (2014). „A comprehensive evaluation of collapsing methods using simulated and real data: excellent annotation of functionality and large sample sizes required“. *Frontiers in Genetics* 5.

Derkach, A, Lawless, JF und Sun, L (2014). „Pooled association tests for rare genetic variants: A review and some new results“. *Statistical Sci* 29.

Dijk, EL van, Auger, H, Jaszczyszyn, Y und Thermes, C (2014). „Ten years of next-generation sequencing technology“. *Trends in genetics: TIG* 30, S. 418–426.

Dolled-Filhart, MP, Lee, M, Ou-yang, Cw, Haraksingh, RR und Lin, JCH (2013). „Computational and Bioinformatics Frameworks for Next-Generation Whole Exome and Genome Sequencing“. *The Scientific World Journal* 2013.

Engle, RF (1984). „Wald, likelihood ratio, and Lagrange multiplier tests in econometrics“. *Handbook of Econometrics*. Hrsg. von Z Griliches† und MD Intriligator. 1. Aufl. Bd. 2. Amsterdam: Elsevier. Kap. 13, S. 775–826. ISBN: 978-0-444-86186-3.

Felix, R, Bodmer, W, Fearnhead, NS, Merwe, L van der, Goldberg, P und Ramesar, RS (2006). „GSTM1 and GSTT1 polymorphisms as modifiers of age at diagnosis of hereditary nonpolyposis colorectal cancer (HNPCC) in a homogeneous cohort of individuals carrying a single predisposing mutation“. *Mutat Res* 602, S. 175–181.

Feng, S, Liu, D, Zhan, X, Wing, MK und Abecasis, GR (2014). „RAREMETAL: Fast and powerful meta-analysis for rare variants“. *Bioinformatics* 30, S. 2828–2829.

Feng, T, Elston, RC und Zhu, X (2011). „Detecting rare and common variants for complex traits: Sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS)“. *Genet Epidemiol* 35, S. 398–409.

Ferrer-Costa, C, Orozco, M und Cruz, X de la (2004). „Sequence-based prediction of pathological mutations“. *Proteins* 57, S. 811–819.

Fischer, G (1983). *Analytische Geometrie*. Wiesbaden: Vieweg. ISBN: 978-3-528-27235-7.

Fisher, RA (1922). „On the interpretation of χ^2 from contingency tables, and the calculation of p“. *J R Stat Soc* 85, S. 87–94.

Fisher, RA (1992). „Statistical methods for research workers“. *Breakthroughs in statistics*. New York: Springer, S. 66–70. ISBN: 978-0-387-94037-3.

Fitze, G, Cramer, J, Ziegler, A, Schierz, M, Schreiber, M, Kuhlisch, E, Roesner, D und Schackert, HK (2002). „Association between c135G/A genotype and RET proto-oncogene germline mutations and phenotype of Hirschsprung's disease“. *Lancet* 359, S. 1200–1205.

Forbes, SA, Beare, D, Gunasekaran, P, Leung, K, Bindal, N, Boutselakis, H, Ding, M, Bamford, S, Cole, C, Ward, S, Kok, CY, Jia, M, De, T, Teague, JW, Stratton, MR, McDermott, U und Campbell, PJ (2015). „COSMIC: Exploring the world's knowledge of somatic mutations in human cancer“. *Nucleic Acids Res* 43, S. D805–D811.

Goeman, JJ und Solari, A (2014). „Multiple hypothesis testing in genomics“. *Stat Med* 33, S. 1946–1978.

Good, P (2000). *Permutation Tests*. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-1-4757-3237-5.

Gorlov, IP, Gorlova, OY, Sunyaev, SR, Spitz, MR und Amos, CI (2008). „Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms“. *Am J Hum Genet* 82, S. 100–112.

Habegger, L, Balasubramanian, S, Chen, DZ, Khurana, E, Sboner, A, Harmanci, A, Rozowsky, J, Clarke, D, Snyder, M und Gerstein, M (2012). „VAT: A computational framework to functionally annotate variants in personal genomes within a cloud-computing environment“. *Bioinformatics* 28, S. 2267–2269.

Han, F und Pan, W (2010). „A data-adaptive sum test for disease association with multiple common or rare variants“. *Hum Hered* 70, S. 42–54.

Hindorff, LA, Sethupathy, P, Junkins, HA, Ramos, EM, Mehta, JP, Collins, FS und Manolio, TA (2009). „Potential etiologic and functional implications of genome-wide association loci for human diseases and traits“. *Proc Natl Acad Sci U S A* 106, S. 9362–9367.

Hoffmann, TJ, Marini, NJ und Witte, JS (2010). „Comprehensive approach to analyzing rare genetic variants“. *PloS one* 5, e13584.

Hotelling, H (1992). „The generalization of Student's ratio“. *Breakthroughs in statistics*. New York: Springer, S. 54–65. ISBN: 978-0-387-94037-3.

Katsanis, SH und Katsanis, N (2013). „Molecular genetic testing and the future of clinical genomics“. *Nat Rev Genet* 14, S. 415–426.

Kimura, M (1968). „Evolutionary rate at the molecular level“. *Nature* 217, S. 624–626.

Kinnamon, DD, Hershberger, RE und Martin, ER (2012). „Reconsidering association testing methods using single-variant test statistics as alternatives to pooling tests for sequence data with rare variants“. *PloS One* 7, e30238.

Knijnenburg, TA, Wessels, LFA, Reinders, MJT und Shmulevich, I (2009). „Fewer permutations, more accurate p-values“. *Bioinformatics* 25, S. i161–168.

Kotowski, IK, Pertsemlidis, A, Luke, A, Cooper, RS, Vega, GL, Cohen, JC und Hobbs, HH (2006). „A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol“. *Am J Hum Genet* 78, S. 410–422.

Kryukov, GV, Pennacchio, LA und Sunyaev, SR (2007). „Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies“. *Am J Hum Genet* 80, S. 727–739.

Ladouceur, M, Dastani, Z, Aulchenko, YS, Greenwood, CMT und Richards, JB (2012). „The empirical power of rare variant association methods: Results from Sanger sequencing in 1,998 individuals“. *PLoS genetics* 8, e1002496.

Lee, S, Wu, MC und Lin, X (2012). „Optimal tests for rare variant effects in sequencing association studies“. *Biostatistics* 13, S. 762–775.

Lee, S, Emond, MJ, Bamshad, MJ, Barnes, KC, Rieder, MJ, Nickerson, DA, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Christiani, DC, Wurfel, MM und Lin, X (2012). „Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies“. *Am J Hum Genet* 91, S. 224–237.

Lee, S, Teslovich, TM, Boehnke, M und Lin, X (2013). „General framework for meta-analysis of rare variants in sequencing association studies“. *Am J Hum Genet* 93, S. 42–53.

Lee, S, Abecasis, GR, Boehnke, M und Lin, X (2014). „Rare-variant association analysis: Study designs and statistical tests“. *Am J Hum Genet* 95, S. 5–23.

Li, B und Leal, SM (2008). „Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data“. *Am J Hum Genet* 83, S. 311–321.

Li, Y, Byrnes, AE und Li, M (2010). „To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests“. *Am J Hum Genet* 87, S. 728–735.

Lin, DY und Tang, ZZ (2011). „A general framework for detecting disease associations with rare variants in sequencing studies“. *Am J Hum Genet* 89, S. 354–367.

Liu, DJ und Leal, SM (2010). „A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions“. *PLoS Genet* 6, e1001156.

Liu, L, Li, Y, Li, S, Hu, N, He, Y, Pong, R, Lin, D, Lu, L und Law, M (2012). „Comparison of next-generation sequencing systems“. *Journal of Biomedicine & Biotechnology* 2012, S. 1–11.

Lumley, T, Brody, J, Dupuis, J und Cupples, A (2012). *Meta-analysis of a rare-variant association test*. Techn. Ber. Stat Tech, University of Auckland.

Luo, L, Boerwinkle, E und Xiong, M (2011). „Association studies for next-generation sequencing“. *Genome Res* 21, S. 1099–1108.

Luo, L, Zhu, Y und Xiong, M (2012). „A novel genome-information content-based statistic for genome-wide association analysis designed for next-generation sequencing data“. *J Comput Biol* 19, S. 731–744.

Luo, L, Zhu, Y und Xiong, M (2013). „Smoothed functional principal component analysis for testing association of the entire allelic spectrum of genetic variation“. *Eur J Hum Genet* 21, S. 217–224.

Ma, C, Boehnke, M, Lee, S und GoT2D Investigators (2015). „Evaluating the calibration and power of three gene-based association tests of rare variants for the X chromosome“. *Genet Epidemiol* 39, S. 499–508.

MacArthur, DG, Balasubramanian, S, Frankish, A, Huang, N, Morris, J, Walter, K, Jostins, L, Habegger, L, Pickrell, JK, Montgomery, SB, Albers, CA, Zhang, ZD, Conrad, DF, Lunter, G, Zheng, H, Ayub, Q, DePristo, MA, Banks, E, Hu, M, Handsaker, RE, Rosenfeld, JA, Fromer, M, Jin, M, Mu, XJ, Khurana, E, Ye, K, Kay, M, Saunders, GI, Suner, MM, Hunt, T, Barnes, IHA, Amid, C, Carvalho-Silva, DR, Bignell, AH, Snow, C, Yngvadottir, B, Bumpstead, S, Cooper, DN, Xue, Y, Romero, IG, 1000 Genomes Project Consortium, Wang, J, Li, Y, Gibbs, RA, McCarroll, SA, Dermitzakis, ET, Pritchard, JK, Barrett, JC, Harrow, J, Hurles, ME, Gerstein, MB und Tyler-Smith, C (2012). „A systematic survey of loss-of-function variants in human protein-coding genes“. *Science* 335, S. 823–828.

Madsen, BE und Browning, SR (2009). „A groupwise association test for rare mutations using a weighted sum statistic“. *PLoS Genet* 5, e1000384.

Maher, B (2008). „Personal genomes: The case of the missing heritability“. *Nature* 456, S. 18–21.

Manolio, TA (2010). „Genomewide association studies and assessment of the risk of disease“. *N Engl J Med* 363, S. 166–176.

Manolio, TA, Collins, FS, Cox, NJ, Goldstein, DB, Hindorff, LA, Hunter, DJ, McCarthy, MI, Ramos, EM, Cardon, LR, Chakravarti, A, Cho, JH, Guttmacher, AE, Kong, A, Kruglyak, L, Mardis, E, Rotimi, CN, Slatkin, M, Valle, D, Whittemore, AS, Boehnke, M, Clark, AG, Eichler, EE, Gibson, G, Haines, JL, Mackay, TFC, McCarroll, SA und Visscher, PM (2009). „Finding the missing heritability of complex diseases“. *Nature* 461, S. 747–753.

Mardis, ER (2013). „Next-generation sequencing platforms“. *Annual Review of Analytical Chemistry (Palo Alto, Calif.)* 6, S. 287–303.

Margulies, M, Egholm, M, Altman, WE, Attiya, S, Bader, JS, Bemben, LA, Berka, J, Braverman, MS, Chen, YJ, Chen, Z, Dewell, SB, Du, L, Fierro, JM, Gomes,

XV, Godwin, BC, He, W, Helgesen, S, Ho, CH, Ho, CH, Irzyk, GP, Jando, SC, Alenquer, MLI, Jarvie, TP, Jirage, KB, Kim, JB, Knight, JR, Lanza, JR, Leamon, JH, Lefkowitz, SM, Lei, M, Li, J, Lohman, KL, Lu, H, Makhijani, VB, McDade, KE, McKenna, MP, Myers, EW, Nickerson, E, Nobile, JR, Plant, R, Puc, BP, Ronan, MT, Roth, GT, Sarkis, GJ, Simons, JF, Simpson, JW, Srinivasan, M, Tartaro, KR, Tomasz, A, Vogt, KA, Volkmer, GA, Wang, SH, Wang, Y, Weiner, MP, Yu, P, Begley, RF und Rothberg, JM (2005). „Genome sequencing in microfabricated high-density picolitre reactors“. *Nature* 437, S. 376–380.

McKusick-Nathans Institute of Genetic Medicine und Johns Hopkins University (Baltimore, MD) (2015). *Online Mendelian Inheritance in Man, OMIM®*. `http://omim.org/`. [Online; Zugriff 12.06.2015].

Mendenhall, W, Sincich, T und Boudreau, NS (1996). *A second course in statistics: regression analysis*. 5. Aufl. New Jersey: Prentice Hall. ISBN: 978-0-13-396821-7.

Metzker, ML (2010). „Sequencing technologies - the next generation“. *Nat Rev Genet* 11, S. 31–46.

Mielke, PW und Berry, KJ (2007). *Permutation methods: A distance function approach*. New York: Springer. ISBN: 978-0-387-69813-7.

Morgenthaler, S und Thilly, WG (2007). „A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST)“. *Mutat Res* 615, S. 28–56.

Morris, AP und Zeggini, E (2010). „An evaluation of statistical approaches to rare variant analysis in genetic association studies“. *Genet Epidemiol* 34, S. 188–193.

Moutsianas, L, Agarwala, V, Fuchsberger, C, Flannick, J, Rivas, MA, Gaulton, KJ, Albers, PK, GoT2D Consortium, McVean, G, Boehnke, M, Altshuler, D und McCarthy, MI (2015). „The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease“. *PLoS genetics* 11, e1005165.

Murken, J (2006). *Humangenetik*. Stuttgart: Georg Thieme Verlag. ISBN: 978-3-13-139297-8.

National Heart, Lung and Blood Institute (NHLBI) (2015). *SeattleSeq Annotation*. `http://snp.gs.washington.edu/SeattleSeqAnnotation/`. [Online; Zugriff 12.06.2015].

Neale, BM, Rivas, MA, Voight, BF, Altshuler, D, Devlin, B, Orho-Melander, M, Kathiresan, S, Purcell, SM, Roeder, K und Daly, MJ (2011). „Testing for an unusual distribution of rare variants“. *PLoS genetics* 7, e1001322.

Nejentsev, S, Walker, N, Riches, D, Egholm, M und Todd, JA (2009). „Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes“. *Science* 324, S. 387–389.

Neyman, J und Scott, E (1966). „On the use of C-Alpha optimal tests of composite hypotheses“. *Bulletin of the International Statistical Institute* 41, S. 477–497.

Ng, PC und Henikoff, S (2003). „SIFT: Predicting amino acid changes that affect protein function“. *Nucleic Acids Res* 31, S. 3812–3814.

Pan, W (2009). „Asymptotic tests of association with multiple SNPs in linkage disequilibrium“. *Genet Epidemiol* 33, S. 497–507.

Pearson, H (2006). „Genetics: What is a gene?“ *Nature* 441, S. 398–401.

Peloso, GM, Rader, DJ, Gabriel, S, Kathiresan, S, Daly, MJ und Neale, BM (2015). „Phenotypic extremes in rare variant study designs“. *Eur J Hum Genet*. Im Druck.

Phipson, B und Smyth, GK (2010). „Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn“. *Stat Appl Genet Mol Biol* 9, Article39.

Pongpanich, M, Neely, ML und Tzeng, JY (2011). „On the aggregation of multimarker information for marker-set and sequencing data analysis: Genotype collapsing vs. similarity Collapsing“. *Frontiers in Genetics* 2, S. 110.

Price, AL, Kryukov, GV, Bakker, PIW de, Purcell, SM, Staples, J, Wei, LJ und Sunyaev, SR (2010). „Pooled association tests for rare variants in exon-resequencing studies“. *Am J Hum Genet* 86, S. 832–838.

Pritchard, JK (2001). „Are rare variants responsible for susceptibility to complex diseases?“ *Am J Hum Genet* 69, S. 124–137.

Pruitt, KD, Tatusova, T und Maglott, DR (2007). „NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins“. *Nucleic Acids Res* 35, S. D61–D65.

Qian, D (2004). „Haplotype sharing correlation analysis using family data: A comparison with family-based association test in the presence of allelic heterogeneity“. *Genet Epidemiol* 27, S. 43–52.

Ramensky, V, Bork, P und Sunyaev, S (2002). „Human non-synonymous SNPs: Server and survey“. *Nucleic Acids Res* 30, S. 3894–3900.

Ramsey, LB, Bruun, GH, Yang, W, Treviño, LR, Vattathil, S, Scheet, P, Cheng, C, Rosner, GL, Giacomini, KM, Fan, Y, Sparreboom, A, Mikkelsen, TS, Corydon, TJ, Pui, CH, Evans, WE und Relling, MV (2012). „Rare versus common variants in

pharmacogenetics: SLCO1B1 variation and methotrexate disposition“. *Genome Res* 22, S. 1–8.

Risch, N und Merikangas, K (1996). „The future of genetic studies of complex human diseases“. *Science* 273, S. 1516–1517.

Risch, N und Zhang, H (1995). „Extreme discordant sib pairs for mapping quantitative trait loci in humans“. *Science* 268, S. 1584–1589.

Rosenbloom, KR, Armstrong, J, Barber, GP, Casper, J, Clawson, H, Diekhans, M, Dreszer, TR, Fujita, PA, Guruvadoo, L, Haeussler, M, Harte, RA, Heitner, S, Hickey, G, Hinrichs, AS, Hubley, R, Karolchik, D, Learned, K, Lee, BT, Li, CH, Miga, KH, Nguyen, N, Paten, B, Raney, BJ, Smit, AFA, Speir, ML, Zweig, AS, Haussler, D, Kuhn, RM und Kent, WJ (2015). „The UCSC Genome Browser database: 2015 update“. *Nucleic Acids Res* 43, S. D670–681.

Schifano, ED, Epstein, MP, Bielak, LF, Jhun, MA, Kardia, SLR, Peyser, PA und Lin, X (2012). „SNP set association analysis for familial data“. *Genet Epidemiol* 36, S. 797–810.

Schork, NJ, Murray, SS, Frazer, KA und Topol, EJ (2009). „Common vs. rare allele hypotheses for complex diseases“. *Current Opinion in Genetics & Development* 19, S. 212–219.

Schwarz, JM, Cooper, DN, Schuelke, M und Seelow, D (2014). „MutationTaster2: Mutation prediction for the deep-sequencing age“. *Nat Methods* 11, S. 361–362.

Seng, KC und Seng, CK (2008). „The success of the genome-wide association approach: A brief story of a long struggle“. *European journal of human genetics: EJHG* 16, S. 554–564.

Sherry, ST, Ward, MH, Kholodov, M, Baker, J, Phan, L, Smigielski, EM und Sirotkin, K (2001). „dbSNP: The NCBI database of genetic variation“. *Nucleic Acids Res* 29, S. 308–311.

Shugart, YY, Zhu, Y, Guo, W und Xiong, M (2012). „Weighted pedigree-based statistics for testing the association of rare variants“. *BMC genomics* 13, S. 667.

Speliotes, EK, Yerges-Armstrong, LM, Wu, J, Hernaez, R, Kim, LJ, Palmer, CD, Gudnason, V, Eiriksdottir, G, Garcia, ME, Launer, LJ, Nalls, MA, Clark, JM, Mitchell, BD, Shuldiner, AR, Butler, JL, Tomas, M, Hoffmann, U, Hwang, SJ, Massaro, JM, O'Donnell, CJ, Sahani, DV, Salomaa, V, Schadt, EE, Schwartz, SM, Siscovick, DS, NASH CRN, GIANT Consortium, MAGIC Investigators, Voight, BF, Carr, JJ, Feitosa, MF, Harris, TB, Fox, CS, Smith, AV, Kao, WHL, Hirschhorn, JN, Borecki, IB und GOLD Consortium (2011). „Genome-wide association analysis

identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits“. *PLoS genetics* 7, e1001324.

Spirin, V, Schmidt, S, Pertsemlidis, A, Cooper, RS, Cohen, JC und Sunyaev, SR (2007). „Common single-nucleotide polymorphisms act in concert to affect plasma levels of high-density lipoprotein cholesterol“. *Am J Hum Genet* 81, S. 1298–1303.

Stitziel, NO, Kiezun, A und Sunyaev, S (2011). „Computational and statistical approaches to analyzing variants identified by exome sequencing“. *Genome Biol* 12, S. 227.

Sun, YV, Sung, YJ, Tintle, N und Ziegler, A (2011). „Identification of genetic association of multiple rare variants using collapsing methods“. *Genet Epidemiol* 35 Suppl 1, S101–106.

Tang, ZZ und Lin, DY (2013). „MASS: Meta-analysis of score statistics for sequencing studies“. *Bioinformatics* 29, S. 1803–1805.

Tang, ZZ und Lin, DY (2015). „Meta-analysis for discovering rare-variant associations: Statistical methods and software programs“. *Am J Hum Genet* 97, S. 35–53.

Treviño, LR, Shimasaki, N, Yang, W, Panetta, JC, Cheng, C, Pei, D, Chan, D, Sparreboom, A, Giacomini, KM, Pui, CH, Evans, WE und Relling, MV (2009). „Germline genetic variation in an organic anion transporter polypeptide associated with methotrexate pharmacokinetics and clinical effects“. *J Clin Oncol.* 27, S. 5972–5978.

Tzeng, JY und Zhang, D (2007). „Haplotype-based association analysis via variance-components score test“. *Am J Hum Genet* 81, S. 927–938.

U.S. Department of Energy und National Institutes of Health (2014). *Human Genome Project*. `http://www.ornl.gov/hgmis`. [Online; Zugriff 12.12.2015].

Valouev, A, Ichikawa, J, Tonthat, T, Stuart, J, Ranade, S, Peckham, H, Zeng, K, Malek, JA, Costa, G, McKernan, K, Sidow, A, Fire, A und Johnson, SM (2008). „A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning“. *Genome Res* 18, S. 1051–1063.

Wang, K, Li, M und Hakonarson, H (2010). „ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data“. *Nucleic Acids Res* 38, e164.

Wang, M und Lin, S (2015). „Detecting associations of rare variants with common diseases: Collapsing or haplotyping?“ *Brief Bioinform* 16, S. 759–768.

Watson, JD und Crick, FH (1953). „Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid“. *Nature* 171, S. 737–738.

Wei, C, Li, M, He, Z, Vsevolozhskaya, O, Schaid, DJ und Lu, Q (2014). „A weighted U-statistic for genetic association analyses of sequencing data“. *Genet Epidemiol* 38, S. 699–708.

Welter, D, MacArthur, J, Morales, J, Burdett, T, Hall, P, Junkins, H, Klemm, A, Flicek, P, Manolio, T, Hindorff, L und Parkinson, H (2014). „The NHGRI GWAS Catalog, a curated resource of SNP-trait associations“. *Nucleic Acids Res* 42, S. D1001–1006.

Wetterstrand, K (2015). *DNA sequencing costs: Data from the NHGRI Genome Sequencing Program (GSP)*. `http://www.genome.gov/sequencingcosts/`. [Online; Zugriff 13.08.2015].

Wilson, EB (1927). „Probable inference, the law of succession, and statistical inference“. *J Am Stat Assoc* 22, S. 209–212.

Wu, B, Pankow, JS und Guan, W (2015). „Sequence kernel association analysis of rare variant set based on the marginal regression model for binary traits“. *Genet Epidemiol* 39, S. 399–405.

Wu, MC, Lee, S, Cai, T, Li, Y, Boehnke, M und Lin, X (2011a). „Rare-variant association testing for sequencing data with the sequence kernel association test“. *Am J Hum Genet* 89, S. 82–93.

Wu, MC, Lee, S, Cai, T, Li, Y, Boehnke, M und Lin, X (2011b). „Rare-variant association testing for sequencing data with the sequence kernel association test“. *Am J Hum Genet* 89, S. 82–93.

Xu, C, Ladouceur, M, Dastani, Z, Richards, JB, Ciampi, A und Greenwood, CMT (2012). „Multiple regression methods show great potential for rare variant association tests“. *PloS One* 7, e41694.

Yandell, M, Huff, C, Hu, H, Singleton, M, Moore, B, Xing, J, Jorde, LB und Reese, MG (2011). „A probabilistic disease-gene finder for personal genomes“. *Genome Res* 21, S. 1529–1542.

Yang, Y, Li, SS, Chien, JW, Andriesen, J und Zhao, LP (2008). „A systematic search for SNPs/haplotypes associated with disease phenotypes using a haplotype-based stepwise procedure“. *BMC Genet* 9, S. 90.

Yue, P, Melamud, E und Moult, J (2006). „SNPs3D: candidate gene and SNP selection for association studies“. *BMC bioinformatics* 7, S. 166.

Zawistowski, M, Gopalakrishnan, S, Ding, J, Li, Y, Grimm, S und Zöllner, S (2010). „Extending rare-variant testing strategies: Analysis of noncoding sequence and imputed genotypes“. *Am J Hum Genet* 87, S. 604–617.

Zelterman, D und Chen, CF (1988). „Homogeneity tests against central-mixture alternatives“. *J Am Stat Assoc* 83, S. 179–182.

Zhang, Q, Irvin, MR, Arnett, DK, Province, MA und Borecki, I (2011). „A data-driven method for identifying rare variants with heterogeneous trait effects“. *Genet Epidemiol* 35, S. 679–685.

Zhu, Q, Ge, D, Maia, JM, Zhu, M, Petrovski, S, Dickson, SP, Heinzen, EL, Shianna, KV und Goldstein, DB (2011). „A genome-wide comparison of the functional properties of rare and common genetic variants in humans“. *Am J Hum Genet* 88, S. 458–468.

Ziegler, A und König, IR (2010). *A statistical approach to genetic epidemiology: Concepts and applications, with an e-learning platform.* 2. Aufl. Weinheim: Wiley-VCH Verlag. ISBN: 978-3-527-32389-0.

Danksagung

Dafür, dass diese Arbeit gelingen konnte, haben viele Personen meinen Dank verdient.

An erster Stelle möchte ich Herrn Univ.-Prof. Dr. A. Ziegler meinen Dank aussprechen. Ihm danke ich für die Vergabe des Themas, für die umfassende wissenschaftliche Betreuung und die zahlreichen hilfreichen Hinweise in der Anfertigung der Dissertationsschrift. Des Weiteren danke ich ihm dafür, dass er mir die Möglichkeit zur Mitarbeit an einer Reihe von spannenden Projekten gegeben hat.

Frau Univ.-Prof. Dr. I.R. König danke ich für die Inspiration zu einigen Konzepten, die diese Arbeit übersichtlicher gestaltet haben. Dr. M.V. Relling danke ich für den Zugang zu den Realdaten und der Möglichkeit zur Veröffentlichung der Ergebnisse aus deren Analyse. Meinen gesamten ehemaligen Kollegen des Instituts für Medizinische Biometrie und Statistik gilt mein Dank für die stets gute Zusammenarbeit, die netten Kaffee-Pausen und das heitere gemeinsame Lachen.

Frau Dr. A. Zech, Herrn Dr. A. Kühnemund und Herrn Dr. A. Schillert danke ich für die zahlreichen Stunden, in denen sie mich bei der Anfertigung der Dissertationsschrift unterstützt haben. Ich danke ihnen insbesondere für die konstruktive Kritik und die gewinnbringenden Hinweise, die diese Arbeit verbessert haben. An dieser Stelle darf der Dank an Herrn C. Schröder, M.Sc. nicht fehlen, der bei vielen kleinen technischen Fragestellungen stets eine große Hilfe war.

Mein Dank gilt auch meiner Familie für die Unterstützung und das Verständnis während des Schreibens dieser Arbeit. Insbesondere möchte ich meiner Mutter A. Sossdorf danken, die mir beigebracht hat, was es heißt nicht aufzugeben.

Lebenslauf

Persönliche Daten

Name	Carmen Dering
Geburtsjahr	1983

Studium

Seit 01/2010	Promotionsstudium an der Universität zu Lübeck
10/2009	Diplom in Wirtschaftsmathematik, Note 1,9 Thema der Diplomarbeit: „Dysons Brownsche Bewegung“
2003–2009	Studium der Wirtschaftsmathematik, Universität Leipzig

Berufliche Tätigkeiten

Seit 11/2015	Wissenschaftliche Mitarbeiterin, Leibniz-Institut für Präventionsforschung und Epidemiologie, Bremen
07/2015–10/2015	Selbstständige Tätigkeit
01/2010–06/2015	Wissenschaftliche Mitarbeiterin, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck
08/2006–12/2009	Werkstudententätigkeit, Comparex (ehemals PC–Ware Information Technologies AG), Leipzig

Publikationsverzeichnis

Zeitschriftenartikel

Dering, C, Schillert, A, König, IR und Ziegler, A (2014). „A comparison of two collapsing methods in different approaches“. *BMC Proc* 8, S8.

Dering, C, Ziegler, A und Hemmelmann, C (2011). „Statistische Analyse von seltenen Varianten: Ein Überblick über Gruppierungsmethoden“. *Biometrie und medizinische Informatik : Greifswalder Seminarberichte, Aachen: Shaker Verlag* Helf 19, S. 31–45.

Schuldt, K, Kretz, CC, Timmann, C, Sievertsen, J, Ehmen, C, Esser, C, Loag, W, Ansong, D, **Dering, C**, Evans, J, Ziegler, A, May, J, Krammer, PH, Agbenyega, T und Horstmann, RD (2011). „A -436C>A polymorphism in the human FAS gene promoter associated with severe childhood malaria“. *PLoS Genet.* 7, e1002066.

Kongressbeiträge[1]

Dering, C, König, IR und Ziegler, A (2014a). „Statistical approaches for gene-based analysis: A comprehensive comparison using Monte-Carlo simulations“. (Poster). International Genetic Epidemiology Society Meeting (IGES), Wien, Österreich.

Dering, C, König, IR und Ziegler, A (2014b). „Statistical approaches for gene-based analysis: A comprehensive comparison using Monte-Carlo simulations“. (Poster). European Mathematical Genetics Meeting (EMGM), Köln.

[1] Vortragender unterstrichen

Dering, C, Nahrstaedt, J, Ziegler, A und Hemmelmann, C (2011). „A review of collapsing methods for the statistical analysis of rare variants“. (Poster). Course in Next Generation Sequencing for rare und common genetic disorders, European School of Genetic Medicine, Bologna, Italien.

Dering, C und Ziegler, A (2015). „Ansätze zur Analyse von seltenen Varianten“. (Vortrag). Biometrisches Kolloquium, Dortmund.

Hemmelmann, C, **Dering, C** und Ziegler, A (2011). „Statistische Analyse von seltenen Varianten: Ein Vergleich von Gruppierungsmethoden“. (Vortrag). Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS), 6. Jahrestagung der Deutschen Gesellschaft für Epidemiologie (DGEpi), Mainz.

Nahrstaedt, J, **Dering, C**, Ziegler, A und Hemmelmann, C (2011). „An overview of collapsing methods for the analysis of rare variants“. (Poster). European Mathematical Genetics Meeting (EMGM), London, Großbritannien.

Ziegler, A, **Dering, C** und Hemmelmann, C (2012). „The Joint Analysis of Rare and Frequent Sequence Variants in Case-Control Studies Using Smoothed Areas“. (Vortrag). International Biometric Conference, Kobe, Japan.